Contents

Introduction 5

Kiswahili as a Low-Resource Language: Sheng and Code Mixing 8

Main Findings 10

1. The Kiswahili Online Public Sphere 10
2. Challenges of Automated Content Moderation in the Kiswahili Language 11
3. The Market for Human Kiswahili Content Moderation 13
4. Global, Local, and Multi-Country Approaches to Content Moderation 15
5. The Trauma of Moderation: Exploitation of Content Moderation Workers 18
    - A. Sugarcoating the Reality of the Job 19
    - B. Silencing Moderators' Voices and Concealing their Identities 20
    - C. Limited psychological support 20
6. A Connected Society of Moderators in East Africa 21

Recommendations 23

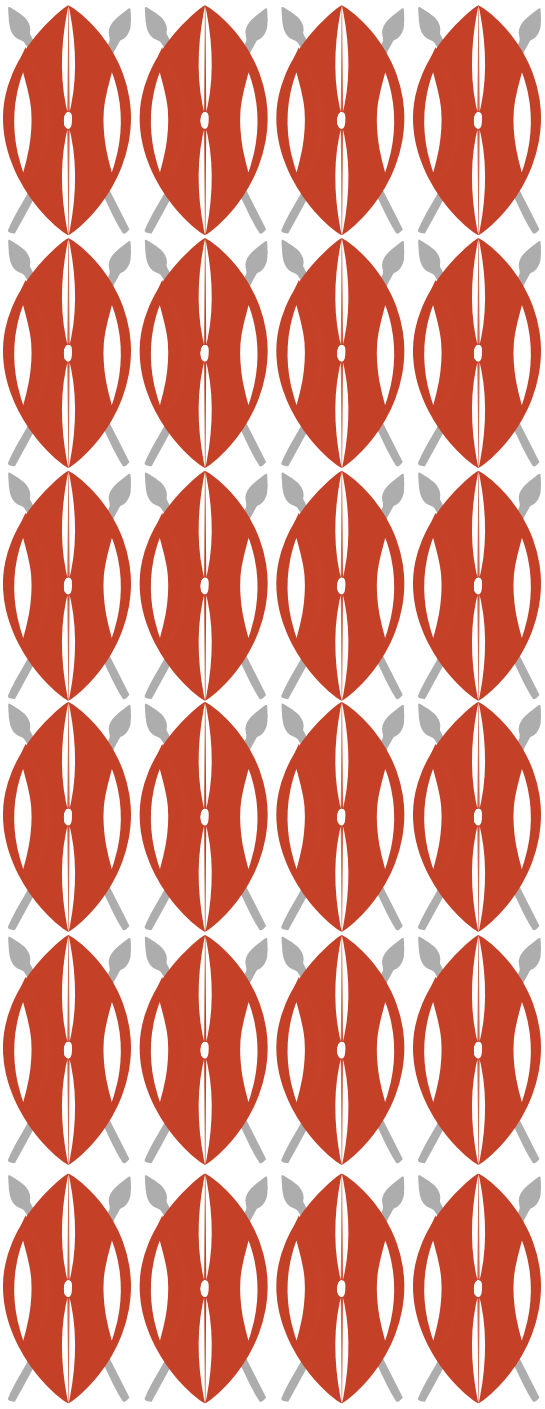
1. Enhance the Diversity of the Content Moderation Teams 23
2. Support and Increase Reliance on Local Kiswahili NLP Researchers 24
3. Clearly Communicate Responsibilities of Moderation Jobs 24
4. Improve Psychological Support for Content Moderators 24

Appendix 26

Methods and Data Collection 26

References 27

Kiswahili as a Low-Resource Language: Sheng and Code Mixing



The multilingualism of East Africa is remarkable. In addition to Kiswahili, Tanzania is home to over 100 indigenous languages, while Kenya has more than 40 languages (Habwe, 2009). Kiswahili is a Bantu language that is spoken by more than 100 million people across East and Central Africa (Shikali & Mokhosi, 2020; Habwe, 2009). Kiswahili is "one of the inner agents of East Africa" and the working language of the East African geo-political area (Habwe, 2009). This language is greatly influenced by Arabic, Hindi, Persian, English, and Portuguese, and it has many geographical variations across East Africa (Githiora, 2002). Some of the same words have different cultural meanings between these variations, which can lead to misinterpretations. For example, "shoga" could mean "friend" in Tanzania but is more likely to mean "homosexual" or "gay" in Kenya. Moderating such variations on social media is challenging.

Many online services are increasingly relying on automated content analysis to moderate user-generated content on their platforms. With the increasing popularity of automation to support content moderation, the "resourcedness" of a language becomes important (Nicholas & Bhatia, 2023). Despite the popularity of Kiswahili, it is considered a low-resource language with limited pre-processed open-source data (Shikali & Mokhosi, 2020). In general, low-resource languages have received less attention in developing training datasets and machine-learning models (Wanjawa et al., 2023). The scarcity of the public domain corpora for low-resource languages, such as Kiswahili, hinders the development of machine learning models, which continue to be optimized primarily for high-resource languages like English (Wanjawa et al., 2023; Wanjawa & Muchemi, 2021).

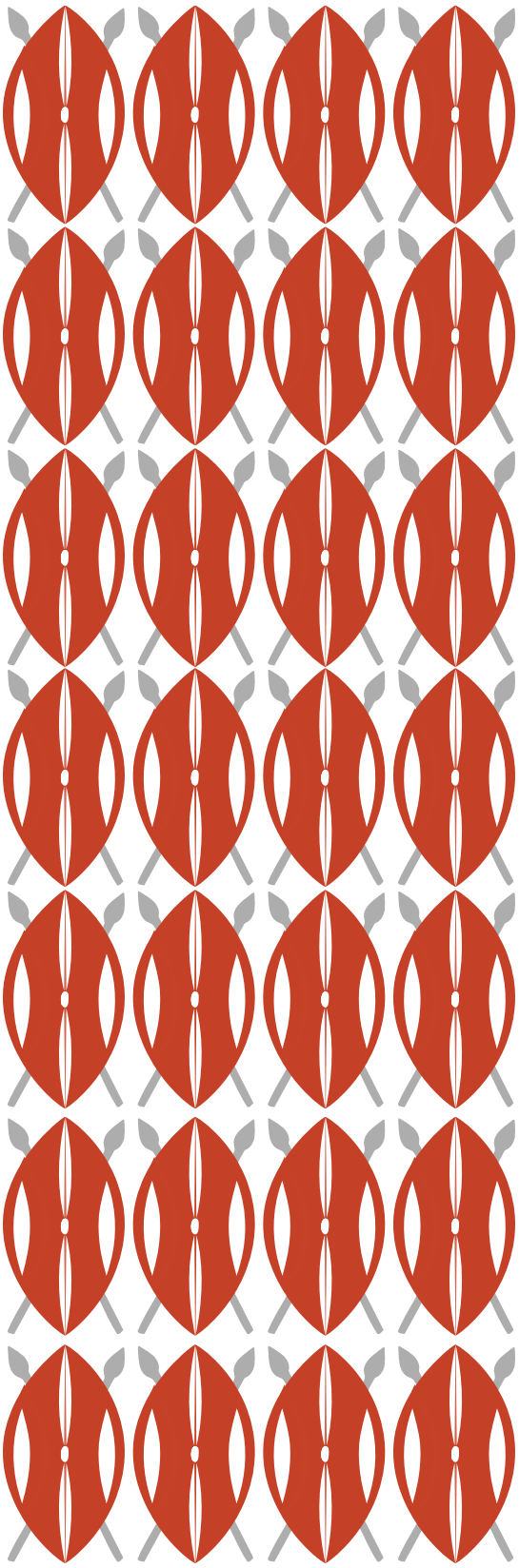
The linguistic evolution of Kiswahili in the post-colonial era has led to the emergence of two linguistically challenging phenomena that further complicate the development of accurate AI models: code-mixing and Sheng. Code-mixing, a term often used interchangeably with code-switching, happens when a speaker incorporates two or more languages within a single sentence or clause (Kanijo, 2017). Code mixing is common among Tanzanian and Kenyan Kiswahili speakers, as many are bilingual and have a working knowledge of English. The blending of the two languages is sometimes referred to as Kiswanglish (Kanijo, 2017).

Sheng, or what is referred to as the youth language, is a variety of Kenyan Kiswahili spoken in informal (*jua kali*) settings outside the classroom and workplace (Githiora, 2018). It is a rapidly evolving language that is closely associated with Nairobi's youth and is the by-product of the region's multilingual dynamics and the search for a modern identity (Githiora, 2018). Sheng originated among young

adults and features a syntax based on Kiswahili, while also incorporating many loanwords from other local languages and foreign languages like English and Hindi (Githiora, 2002).

Given the variations, complexities, and scarcity of training datasets, the current content moderation systems used to regulate Kiswahili content online have significant shortcomings that affect Kiswahili speakers. This report examines the Kiswahili online public sphere, the content moderation market, and content moderation approaches. This report also proposes recommendations on how to improve Kiswahili content moderation.

Main Findings



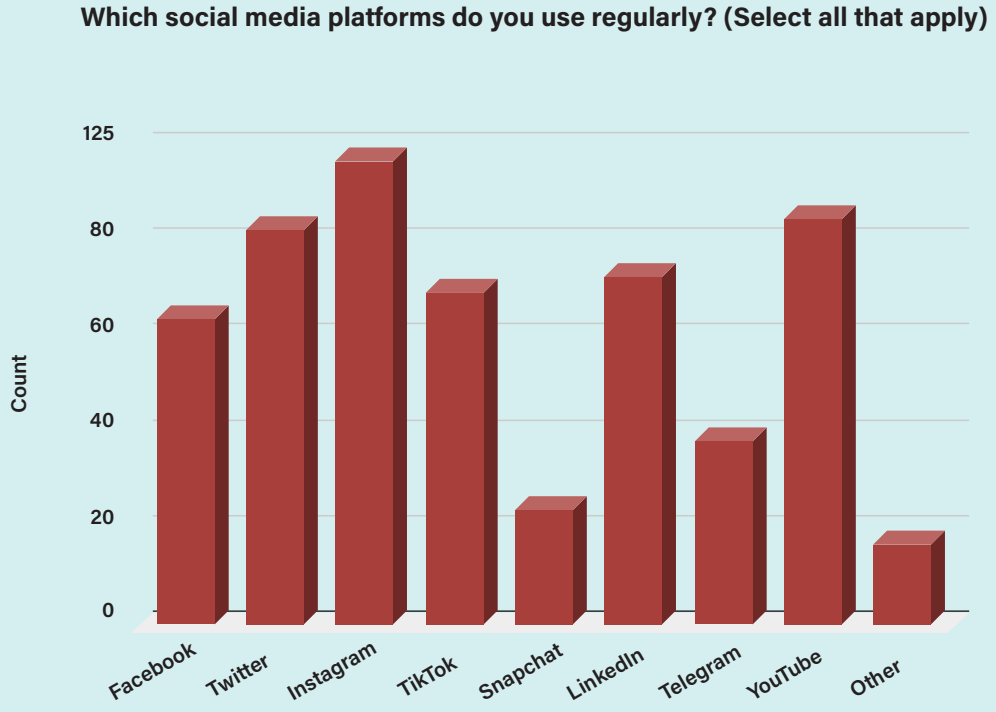
1. The Kiswahili Online Public Sphere

Based on survey results gathered from 143 participants who are frequent social media users in Kenya and Tanzania, we found that Instagram, YouTube, and X (formerly known as Twitter) are the most popular social media platforms. TikTok has surpassed Facebook in popularity among our Swahili participants, underscoring the evolving digital habits of people in the region and the increasing influence of visual media on their social interactions (see Figure 1). That may be a reflection of a broader trend, as TikTok is the most downloaded application in both Kenya and Tanzania (Madung, 2022; Appfigures, 2024).

Participants were concerned about the spread of hate speech and misinformation in Kiswahili. **More than 78% of our survey participants shared their concerns about the spread of misleading content and inciting materials online.** Previous research revealed that misleading, polarizing, and inciting content rapidly spread on TikTok and Twitter in Kenya and enabled bad actors to spread malicious content online (Madung, 2022a; Madung, 2022b).

Moreover, about 70% of participants indicated that they have at least once reported content that they believed was violating a platform's policy. Among those participants, about 46% indicated that their reports were reviewed and action was taken accordingly (see Figure 2). Results show that about 37% were either satisfied or very satisfied with the current reporting mechanisms used by social media companies. Our interview participants talked about relying on third-party escalation and civil society groups to report harmful content in some cases. A few content creators indicated that they also utilized their personal connections to internally escalate their concerns. A Tanzanian content creator who tried to report a Facebook account that impersonated one of her relatives said:

"So I did several reports on the account, but it was still there and not taken down. Because the account was in Swahili, there was no way it would be removed. I don't know, maybe the review was taking time. After a month, I had to communicate with someone from Meta, telling him what was happening and that I had reported it, but nothing had been done. He asked me to send an email to his official email. When I sent the email, it took less than two hours, and the account was taken down. So it means if I didn't have any contact from a Meta person, there was nothing I could do." (Content creator, Tanzania, July 2024)



▲ **Figure 1.** Regularly used social media among Kiswahili speakers (n=143. Participants were allowed to select multiple options). *Source* CDT's online survey (July-Sep 2024).

Additionally, the phenomenon of "mass reporting" was observed in the Kiswahili context, though it was not a common tactic among participants. Interview participants expressed a tendency to rely on their social capital to address content they were unable to remove through official reporting mechanisms. One content creator noted,

> *"I also reported content in 2020 when I was just coming from university, and someone opened an account using my name, posting my pictures with really bad captions. Then when I reported it, like nothing happened, and I had to invite my friends, like, 'Hey, please help me to report this.' But it took really, really, really a long time for them to remove the account." (Content creator, Tanzania, July 2024)*

2. Challenges of Automated Content Moderation in the Kiswahili Language

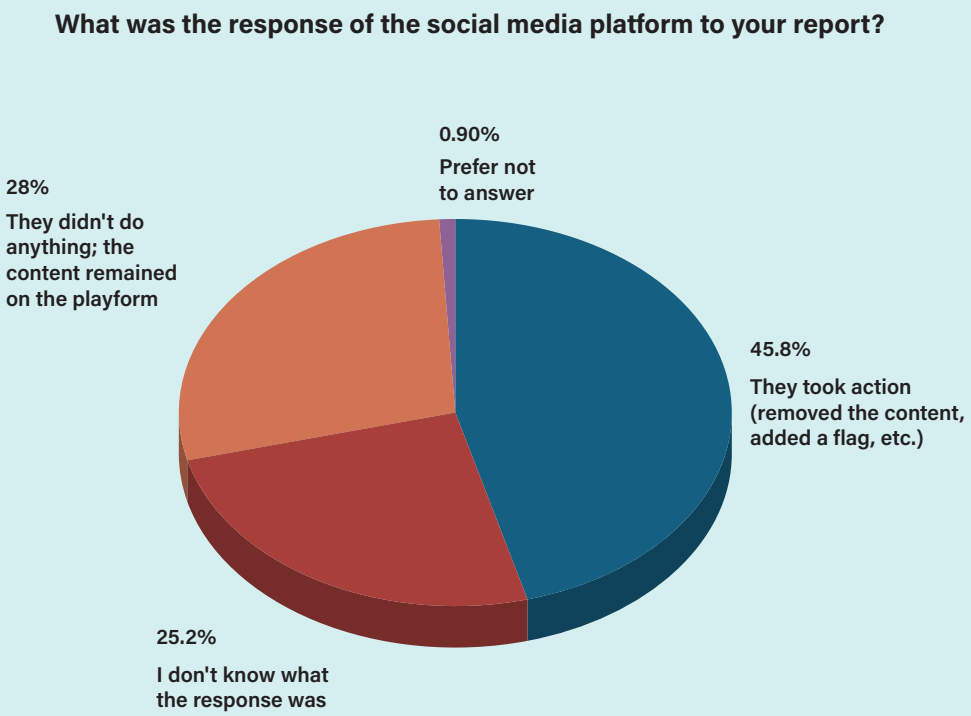
Tech companies increasingly use a combination of classifiers and automated tools to moderate non-English content and rely on multilingual language models to moderate and detect harmful content for low-resource languages, to compensate for the lack of digitized

training data in these languages (Nicholas & Bhatia, 2023). Data for low-resource languages is frequently of low quality, often suffering from mistranslations or sourced from a handful of specific sources like religious texts and Wikipedia. Multilingual language models seek to bridge these data gaps by applying semantic rules inferred from higher-resource languages onto lower-resource ones, but they typically depend on English texts, which can lead to the introduction of unsuitable values and assumptions into other languages (Nicholas & Bhatia, 2023). Better datasets are needed for Kiswahili and other low-resource languages to have more accurate automated content moderation. However, collecting and annotating such datasets has its limitations.

We conducted a roundtable attended by ten Kiswahili Natural Language Processing (NLP) and linguistic researchers from Kenya, Uganda, and Tanzania. The discussions revealed significant linguistic complexities, including the evolution of Kiswahili, Sheng, and the integration of Code-Mixing with local dialects. These factors complicate data collection and model generalization, especially given the diversity in cultures across East Africa. Sheng is a rapidly evolving language, with new words and phrases emerging from trending songs, social media memes, and sociopolitical events. Variations of Sheng can also differ significantly between urban, peri-urban, and densely populated areas, presenting unique challenges for NLP models compared to more stable dialects.

In addition to the linguistic challenges, the process of data collection and annotation is also impacting the quality of models. Participants identified critical data access issues that led them to rely on community-based resources to collect training data. Data annotation is also costly, and participants indicated that they relied on students or friends, while some others indicated that they turned to AI tools like ChatGPT for annotation. Furthermore, finding annotators familiar with Sheng is particularly difficult, given the rapidly evolving nature of this language.

In addressing these challenges, participants indicated that they rely on various tactics. They emphasized that collaborating with local communities is essential to enhance data accuracy and reliability. Participants also addressed the need to develop task-specific models trained on comprehensive Kiswahili datasets that also incorporate Sheng. Finally, participants advocated for increased funding and institutional support to build capacity in AI and linguistic research, underscoring the need for technology companies to engage with local Natural Language Processing (NLP) experts in East Africa to enhance the effectiveness of automated content moderation for the Kiswahili language.



▲ **Figure 2.** The response of social media platforms to participants' reports (% of survey participants who stated that they have previously reported content on social media, n=107). *Source* CDT's online survey (July-Sep 2024).

3. The Market for Human Kiswahili Content Moderation

In addition to automation, human content moderation is essential for managing Kiswahili content. East Africa attracts numerous tech companies looking to outsource the demanding task of content moderation to specialized firms, particularly in Kenya. The country's young workers, who are willing to undertake challenging work for low wages, along with its multilingualism featuring a clear English accent, supportive legal frameworks, and relatively flexible immigration policies, have made Kenya an appealing destination for outsourcing content moderation labor. A Kenyan digital rights advocate stated:

> *"So it's a set of different factors that made Kenya a choice for them [social media companies]. It's really about having a young population; immigration-wise, it's easier to bring people from different African countries to work in Kenya than it is to get Kenyans to work in a different African country. And then there's the language aspect as well." (Digital rights advocate, Kenya, Aug. 2024)*

During this study, we identified two primary content moderation vendors that provide or recently provided services for Kiswahili content: Sama (formerly known as Samasource) and Majorel (which has been acquired by Teleperformance). Both companies are based in Nairobi.

Sama, which originated in California, identifies itself as an “ethical AI” outsourcing company with core expertise in data annotation (Perrigo, 2022). According to its website (Sama, n.d), content moderation represented only 4% of its business, and the rest focused on data annotation. According to our interview participants, Sama began its moderation contract with Meta in 2019 and discontinued it in January 2023 after receiving international backlash due to the exploitation of moderators, leaving more than 180 moderators unemployed (Ambrose, 2023; Bhalla, 2023). Consequently, content moderators filed a lawsuit against Sama and Meta in April 2023, seeking compensation for what they termed an "unconstitutional layoff." The CEO of Sama later expressed regret over taking on Meta’s content moderation responsibilities, announcing that the company would no longer accept content moderation work and would instead focus solely on data labeling for Meta and other clients. (Ambrose, 2023; Bhalla, 2023). According to interview participants, Meta has moved its content moderation operation out of Kenya and relocated its moderation labor to Ghana, although this has not been officially confirmed.

The second company leading content moderation for Kiswahili content in East Africa is Majorel. According to our interview participants, TikTok contracted with Majorel in 2023 to moderate content for the Sub-Saharan region, including the Kiswahili market. Additionally, after Sama ended its contract with Meta, the latter contracted with Majorel to assume the content moderation contract; however, this arrangement did not last long. Upon taking over, Majorel did not hire the moderators laid off by Sama, opting instead to recruit new employees to handle the content moderation for Meta and TikTok. Consequently, the former Sama moderators included Majorel in their lawsuit for denying them employment opportunities (Kannampilly & Malalo, 2024; Agbetiloye, 2024). To this date, Majoral is contracted to handle TikTok content moderation.

While the Kiswahili content moderation market has been primarily associated with these two companies, some participants suggest that other firms may be quietly contracting with tech companies. This reality remains obscured from public view due to a culture of secrecy surrounding content moderation work. Originally, this secrecy was established to prevent “unscrupulous users” from learning about

content moderation policies and discovering ways to circumvent them (Roberts, 2019). However, secrecy has since been overused and developed into a widespread culture within content moderation practices. One participant, who is a legal expert, expressed this concern:

> *"The entire context of the model with which big tech companies work is privacy and a lot of secrecy around their operations. The only ones that people became publicly aware of are Sama and Majorel because of the court cases, but that's not to say they are the only companies doing content moderation work." (Lawyer, Kenya, Aug 2024)*

4. Global, Local, and Multi-Country Approaches to Content Moderation

In general, we identified three approaches to moderating Kiswahili content: global, local, and multi-country. Moderators working in the Kiswahili market are assigned content in two primary languages: English and Kiswahili. However, the moderation approaches vary in terms of (a) the number of countries from which the reviewed content originates and (b) whether policy implementation varies from one country to another. In our first report in this series, we discussed global and local approaches to content moderation; for this report, we will re-examine those within the Kiswahili context and explore the multi-country approach identified in the Kiswahili market.

Global content moderation is primarily employed by US-based social media companies (Elswah, 2024). In this approach, Kiswahili content moderators applied the same policies for all users. **Moreover, content moderators who reviewed Meta content were asked to moderate English-language content from all over the world.** A former Kiswahili content moderator expressed:

> *"We would moderate the tickets from Norway, moderate tickets from the US, moderate tickets from Tanzania. We'd moderate tickets from South Africa. We'd moderate tickets from Zimbabwe— I mean everywhere, everywhere. So yes. As long as they were in English and at the top it was written that the market was from Norway, you just had to action the ticket. You couldn't leave the ticket on the platform. So yeah, which was our target, and we had to make sure that we actioned the ticket." (Content moderator, Kenya, Aug. 2024)*

Localized content moderation entails tailoring some policies, especially those related to cultural matters, to meet the local norms and laws in a certain country (Elswah, 2024). This approach is exemplified by TikTok. A content moderator said:

> *"Normally there is an exemption for culturally appropriate content. For example, if something is practiced in a certain place, there is a policy to cover that. We sometimes get videos from certain communities in Kenya whereby the ladies are maybe nude and have applied some red foundations during a cultural setting function; such content is allowed ... For instance, we have people who normally have cultural festival days, and you find the women dancing, they have applied red ochres and they're not covering their breasts. However, this is a cultural setting." (Content moderator, Kenya, Aug. 2024)*

Another moderator who reviewed content for TikTok talked about the variations in the enforcement of the drug policy by TikTok in different East African countries, noting:

> *"So, in Tanzania, we have this drug called Khat. It's a type of leaf that people chew for pleasure. At the beginning of this project, we were not allowed to tag Khat in Kenya. However, when you see content from Tanzania, you must tag it because Tanzania does not accept that drug. They don't use it. So when you see content from Tanzania that includes Khat, you will tag it as a drug. In Kenya, though, it is considered a cash crop, so it is allowed." (Content moderator, Kenya, 2024)*

Another factor we found in this approach is that moderators only reviewed English or Kiswahili local content coming from East Africa. **Unlike the global content moderation approach, moderators applying the localized approach did not have to assess English-language content from Western countries.**

Although the local approach to content moderation attempts to address the regional and linguistic context of content, limited diversity in vendor employment can significantly produce some of the same problems found in the global approach. **Vendors specializing in Kiswahili content moderation have predominantly concentrated their hiring efforts on moderators from Kenya, with only a handful of individuals who have some family ties in Tanzania.** This narrow hiring approach may limit the understanding of the diverse

dialects and cultural contexts within the Kiswahili-speaking countries, potentially overlooking nuances beyond the more widely recognized differences. This lack of representation in the moderation teams may not only impact the accuracy of content moderation in some cases but also reflect a broader disregard for the linguistic and cultural nuances that are essential for accurate and sensitive content handling. A former moderator noted:

> *"If these people thought Swahili market is also important, they would have brought even at least two people from Tanzania to help with the Kenyan, you know. So if they had maybe brought in, since the content moderation was in Nairobi, right? So they just assumed that, 'Swahili is just Swahili, so let's just work with Kenyans only,' you know. There were no Tanzanians." (Content Moderator, Kenya, Aug. 2024)*

This unrepresentative hiring of Kiswahili moderators results in significant issues in moderation, stemming from a lack of understanding of the diverse cultures, dialects, and contextual nuances prevalent in various countries across East Africa. In many instances, when moderators from Kenya received content originating from Tanzania, they had to rely upon their own best judgment or ask for assistance from colleagues or quality analysts. Moderators sought out help because of concerns that failing to complete their tasks accurately would negatively impact their performance metrics. A former content moderator stated:

> *"Coming to Tanzanian language, that's where the challenge lies, 'cause they have their type of communicating. If you don't understand the meaning, you end up losing your action. What would happen is you would do the job. You would try to assume because the quality analyst was also Kenyan. You would try to ask him or her, then they would be like, 'Well, just action it and then we'll flag it during training, then we discuss it.' But at the same time, if you action it and then you find it is wrong, you are scored down again. ... we had to maneuver to understand the language over time, grasp the meaning, and understand how they communicate in Tanzanian Swahili, especially compared to Kenyan Swahili and Congolese." (Former content moderator, Kenya, Aug. 2024)*

The third approach is "one language and multi-country," which goes beyond language-based moderation to focus on moderating content according to each country's norms, context, and culture. A notable example of this approach is the Tanzanian platform "JamiiForums," established in 2006 originally with the name JamboForums. According

to one member of JamiiForums, the platform's aim is to serve Swahili users who come from different cultures and contexts. The platform's leadership recognizes that Kiswahili has multiple variations and access to the local contexts where these language varieties exist is vital to the human rights mission of the firm. By prioritizing local context, this platform demonstrates a commitment to preserving freedom of expression and cultural relevance in its moderation practices. One JamiiForums team member said:

> *"We have people from Rwanda who review content from Rwanda because they know the local context. This is why I believe Meta gets it totally wrong whenever they remove something that was supposed to stay in the public domain and for JamiiForums, many state actors demand us to remove content, and we respond to them that this is going to remain because we know the local contexts and we know why they are reporting such content." (JamiiForums member, Tanzania, Aug. 2024)*

This approach, however, relies on more human moderators and less automation. JamiiForums utilizes automation in two limited capacities: (a) removing spam and insults and (b) removing content that has been reported multiple times. In the latter case, the content is reviewed by a moderator after being removed from the public domain, with the moderator making the final decision.

Despite being uncommon, and costly in its implementation, this method of content moderation has gained a positive reputation among digital rights activists. Many participants we spoke with praised the efficacy of content moderation implemented by JamiiForums, highlighting how it has achieved representative and fair moderation of Kiswahili content. However, the JamiiForum approach has a user base of approximately three million users per day, raising questions about whether other digital platforms with much larger user bases would be able to apply a similar strategy.

5. The Trauma of Moderation: Exploitation of Content Moderation Workers

Content moderation in East Africa has garnered global attention as moderators shared their traumatizing experiences, from reviewing extremely graphic material to enduring long working hours. The findings in this report reveal a systematic exploitation of content

moderation workers, particularly those in the Kiswahili market. This exploitation takes various forms and manifestations, and here we highlight only three.

A. SUGARCOATING THE REALITY OF THE JOB

Content moderation vendors tend to sugarcoat moderation jobs by concealing the nature of the role and the graphic content employees will review. During job interviews, hiring personnel often did not provide sufficient information about the role to the participants, many of whom were eager to start any lawful job that offered adequate pay. In some instances, job applicants misinterpreted the role and what was expected of them. Some of the participants we interviewed for this study thought that they were going to work as content "creators" rather than moderators or that they were going to be employed directly by Meta or TikTok. A former content moderator who worked at Sama stated:

> *"So during the interview, they did not even tell us what we'll be doing exactly. So when they were interviewing me, they were asking normal questions, where do you see yourself in the next five years? Things like that. But now, in the middle, they will ask random questions. You'll wonder, why are they asking? But you'll be like, 'Ah, I just need this job.' ... But then I passed the interview, I got a job, we went for training. I still did not actually know clearly what we were doing. So the content moderation thing was all new to all of us." (Content Moderator, Kenya, Aug. 2024)*

During the training, participants were introduced to the role and their expected responsibilities for the first time. However, the severity of the content was downplayed, and the examples provided during the training were less graphic than what they often encountered on the job. A former content moderator stated:

> *"The training was nothing compared to the job. Actually, the training was easier. And there were no graphic contents on training until we got the actual job now, the production part... They kept saying we'll see graphic content, but the images they were showing to trainers were not as graphic as the images we were getting or the videos we were getting when we were doing the actual job after training. I was shocked. I was just shocked.*

And I'm like, 'Oh, God, I need another job.' But I also, you know, getting employment is not that easy. And also the money was a bit good. Yeah. So, I was like, 'Where will I get a job like this?' So I was just there due to a lack of employment out here." (Content moderator, Kenya, Aug. 2024)

B. SILENCING MODERATORS' VOICES AND CONCEALING THEIR IDENTITIES

Due to the culture of secrecy in the content moderation business, interview participants discussed how **moderators were not allowed to share the policies or the content they review as well as any details about the nature of their work with their families and close friends. This prevented them from sharing their trauma, further exacerbating their distressing experiences.** This lack of openness not only isolated them but also deprived moderators of the emotional support they saw as essential for coping with the challenges of their roles. A former content moderator stated:

"I'm an IT agent. If anyone would ask me what type of work I do I will tell them I'm an IT agent. We were not allowed to mention that we are content moderators or even where our offices were." (Content Moderator, Aug. 2024)

C. LIMITED PSYCHOLOGICAL SUPPORT

The demands of content moderation often create a challenging environment for moderators, who feel pressured to prioritize work obligations over their mental health. The emphasis on meeting key performance indicators (KPIs) contributes to a culture that fails to recognize the emotional toll of the job. Many participants spoke about their inability to seek counseling support because they wanted to "action more tickets." Additionally, interview participants noted that they were unable to take an additional break after encountering traumatic content because they had exhausted their one-hour break. When requests for wellbeing breaks were denied, moderators had to cope with emotional burdens without any additional support. Participants indicated that preventing them from wellbeing breaks prohibited them from processing trauma, leaving them to navigate their feelings in isolation. A participant noted:

"When you wanted time off, they would tell you, you know, you've exhausted your breaks, you cannot go anywhere. And sometimes you've seen something that is like, you know, that is weighing you down, that is affecting you so bad and they're like,

> *nah, you have already exhausted your breaks. We used to have a one-hour break only. So that break is the one that you'll distribute on your break, lunch, your bathroom breaks. They're like, 'Nah, you cannot do that. You've exhausted your break.' So you have to sit there with your emotions, with your everything." (Content moderator, Kenya, Aug. 2024)*

Additionally, participants talked about their prolonged exposure to traumatic content and the many profound ways it affected them. In particular, they expressed that it altered their perceptions of the world, disrupting their social responsibilities. This impact is particularly significant for those who are breadwinners, as their experiences can affect not only their own lives, but also the well-being of their families and communities. A digital rights advocate stated:

> *"It's not just that they [moderators] are suffering from, you know, a little bit of debilitating sickness; it's that it affects how they perceive life. When they consume extremist content or when they consume a lot of violent content, it changes how they view society, the people around them. There are people who've not been able to care for their children after that, because you've been exposed to a lot of child abuse. So what ends up being is that it changes all their dynamics in life. Their social interactions, their work, even their ability to hold on to work, to get work, and to hold on to it is affected. And that really impacts their quality of life." (A digital rights advocate, Kenya, Aug. 2024)*

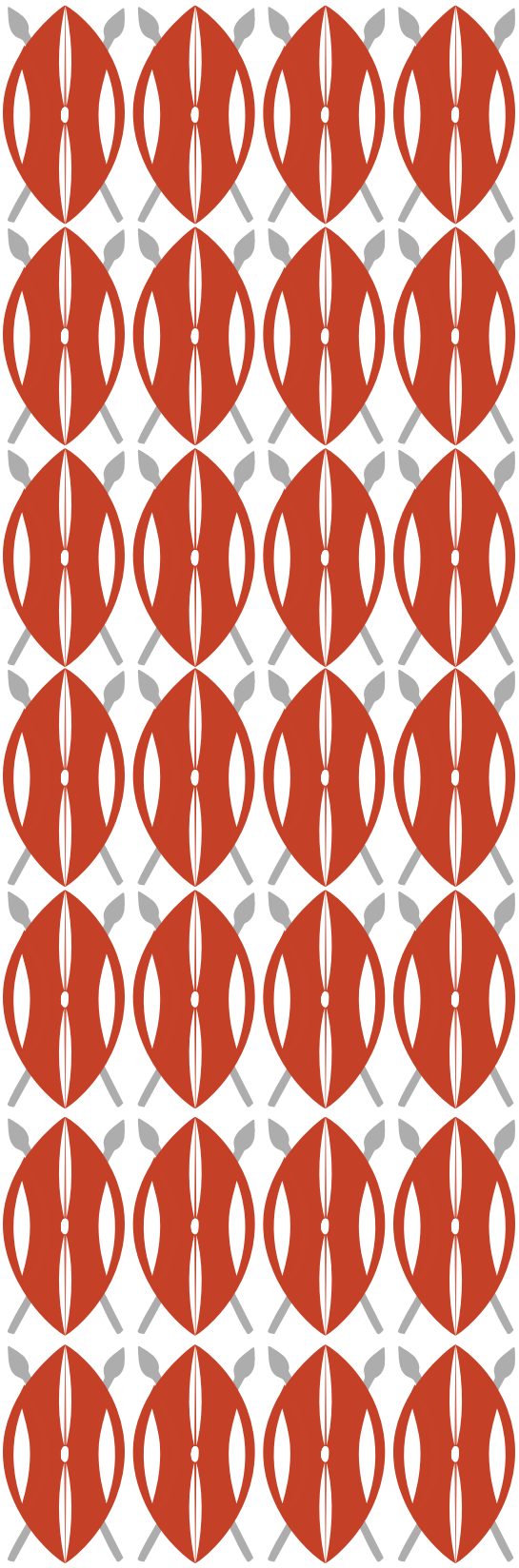
6. A Connected Society of Moderators in East Africa

In the face of these seemingly repressive characteristics of their jobs, Kiswahili content moderators have resisted by uniting to advocate for their rights and those of their peers. Moderators in Kenya, despite the kind of secrecy imposed on them, meet regularly, form online chat groups, and exchange ideas. A former content moderator noted:

> *"We actually sometimes meet in seminars. We have groups that we created, so that at least everyone is aware of whatever is happening. We joined a union that is actually trying to help us achieve our aim. That is making sure that content moderation jobs, people are not taken advantage of." (Content Moderator, Kenya, July 2024)*

Content moderators in Kenya formed the first African Content Moderators Union, with more than 150 members (Perrigo, 2023). The formation of this union has fostered a sense of solidarity and collective advocacy within the community. The union's primary objective is to ensure that content moderation professionals are not exploited, reflecting a commitment to improving working conditions and promoting fair treatment within the industry.

Recommendations



Content moderation is a complex puzzle, and moderators are a crucial piece. When they are exploited or lack adequate support, moderators are likely to make inaccurate judgments affecting the ability of Kiswahili speakers to meaningfully use online services. This is particularly evident when moderators are tasked with reviewing content from unfamiliar cultures and contexts, leading to inconsistencies that can undermine the integrity of their moderation efforts. To improve the moderation of Kiswahili content, it is essential to address these challenges and prioritize the well-being of moderators. We offer the following recommendations that can be applied not only to the Kiswahili market but likely to all languages in the Global South.

1. Enhance the Diversity of the Content Moderation Teams

In the Kiswahili context, we found that only Kenyans were hired to work with the Nairobi-based vendors. These Kenyan moderators were tasked with reviewing content from across East Africa, which is rich with diverse contexts. This limitation led to errors in moderation, as the moderators struggled to assess non-Kenyan content, often presented in different Kiswahili dialects. Tech companies should ensure that their content moderation vendors hire a diverse team of moderators from various countries. Additionally, vendors should facilitate the immigration and relocation of moderators to effectively cover the full spectrum of language variations.

Moderators should also be assigned a more reasonable workload. Those we spoke with reported reviewing hundreds of tickets daily, with one moderator mentioning a daily goal of one thousand tickets. Such a high volume forces moderators to make decisions quickly, often leading them to view only a fraction of the content before reaching a conclusion. In situations where the content is unfamiliar and requires further investigation, they lack the necessary time, resulting in inaccurate decisions. Tech companies need to require vendors to hire additional moderators from diverse backgrounds, thereby improving moderation quality and reducing the workload and the burden for Kiswahili moderators and others who review content in Global South languages.

2. Support and Increase Reliance on Local Kiswahili NLP Researchers

The presence of code-mixing, Sheng, and numerous variations of the Kiswahili language presents significant challenges for fair and accurate automated content moderation. However, many local experts in the East African region have developed and curated Kiswahili datasets reflecting these diverse variations. They have relied on limited resources and innovative data collection methods to create and process these datasets. Tech companies should leverage the expertise of these local professionals, utilizing their invaluable knowledge to develop accurate AI models not only for the Kiswahili language but also for over 2,000 other African languages.

3. Clearly Communicate Responsibilities of Moderation Jobs

All of the moderators we spoke with reported experiencing trauma as a result of the graphic content they encountered in their work. It is crucial that moderators are informed about the nature of the job early in the hiring process. They should be made aware that they will be reviewing graphic and harmful content daily, and that this exposure may lead to trauma. Providing potential candidates with a fuller description is essential because it affords them the ability to make an informed choice about whether to take the job. During the training stage, candidates should also be told more about the graphic content they will be moderating and how to deal with it emotionally.

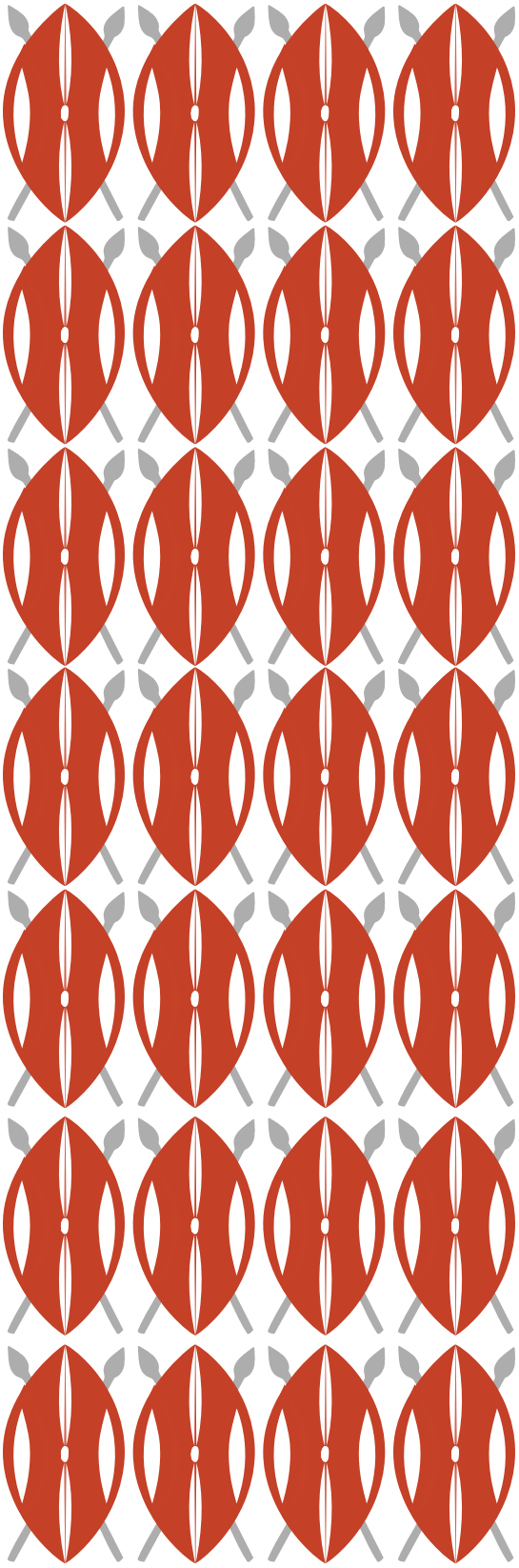
4. Improve Psychological Support for Content Moderators

Moderators should be given access to well-being breaks whenever they need them in addition to their regular one-hour daily break. Denying them these essential mental breaks is unacceptable. Also, vendors should hire more counselors who are specifically trained to understand the nature of content moderation and the impact of reviewing graphic content. Accessing counselors should not negatively impact moderators' performance. Instead, individual or group counseling sessions should be prioritized. Moderators should feel encouraged to seek counseling rather than being discouraged by potential deductions in their KPIs and salaries.

Moderators should be celebrated, appreciated, and acknowledged for their role in making the online space safer by removing harmful content. Describing themselves as the "Internet police," moderators recognize the importance of their work in protecting "their people" and families from exposure to harmful material. However, they often feel unappreciated, unable to be proud of their efforts to shield others from traumatic content. This lack of recognition prevents them from celebrating their achievements or sharing their emotional struggles.

The secrecy surrounding this job must be addressed to ensure moderators have the support they deserve. Fostering mutual trust between companies and moderators could lead to better well-being and more accurate decision-making. One participant suggested establishing an “International Moderators Day,” which could be a valuable way to acknowledge and appreciate their contributions rather than treating them like undercover Internet cops.

Appendix



Methods and Data Collection

To study content moderation systems for the Kiswahili language, we used a mixed-method approach, combining qualitative and quantitative methods. First, we conducted semi-structured, in-depth interviews to assess the information ecosystem in East Africa, the governance of Kiswahili by tech companies, and the challenges digital rights activists face in advocating for equitable moderation for all the different variations of the Kiswahili language. We interviewed 23 participants from a variety of backgrounds, 14 of which were current or former Kiswahili content moderators from third-party vendors in Kenya. We also interviewed four digital rights advocates and lawyers from East Africa and four content creators and influencers from Kenya and Tanzania who publish content in Kiswahili. Lastly, we interviewed one participant who is currently working in a US-based social media platform.

All interviews were conducted online. The interviews occurred between July and September 2024 and were predominantly conducted in English. Field notes were taken during the interviews, and the interview recordings were later transcribed and analyzed to find common themes among the participants.

We also conducted an online survey to understand how frequent social media Kiswahili users in Kenya and Tanzania use online services, report inappropriate content, and deal with content removals. The survey asked questions about users' trust in social media companies and their perceptions of the toxicity of online platforms in their region. The Alchemer platform was used to distribute the online survey from August 4 to September 5th, 2024. A modest honorarium of US$10 was offered for participation. The survey was administered in English and Kiswahili. We were able to collect a sample from 143 participants from Kenya and Tanzania. About 33% of the participants were from Tanzania and 66% from Kenya. We had a gender-balanced sample of participants, with about 71 self-identified as males, 71 as females, and one non-binary.

We also organized a roundtable discussion with Natural Language Processing (NLP) researchers from East Africa who work on the Kiswahili language. During the roundtable, we addressed the challenges they face in data collection, annotation, and processing.

References

Agbetiloye, A. (2024). Meta faces legal action in Kenya over layoffs of content moderators. *Business Insider Africa.* https://africa.businessinsider.com/local/markets/meta-faces-legal-action-in-kenya-over-layoffs-of-content-moderators/kkc5g8p [perma.cc/H63L-963W]

Ambrose, T. (2023, August 16). CEO regrets her firm took on Facebook moderation work after staff 'traumatised.' *The Guardian.* https://www.theguardian.com/technology/2023/aug/16/sama-ceo-regrets-firm-took-on-facebook-moderation-work-kenya-staff-allege-exposure-graphic-content [perma.cc/2JHD-JZZS]

AppFigures. (2024). *Top Apps & Games for Android on Google Play in Kenya.* Appfigures. https://appfigures.com/top-apps/google-play/tanzania/top-overall [perma.cc/698B-MPE4]

Bhalla, N. (2023, January 24). Meta urged to boost Africa content moderation as contractor quits. *Reuters.* https://www.reuters.com/world/africa/meta-urged-boost-africa-content-moderation-contractor-quits-2023-01-24/ [perma.cc/GFF5-J2PB]

De Gregorio, G., & Stremlau, N. (2023). *Inequalities and content moderation. Global Policy,* 14(5), 870–879. https://doi.org/10.1111/1758-5899.13243 [perma.cc/WQ4X-6SHX]

Dean, M. (n.d.). *Contracts of silence.* Columbia Journalism Review. Retrieved October 4, 2024, from https://www.cjr.org/special_report/nda-agreement.php/ [perma.cc/FJK3-UPFU]

Dzahene-Quarshie, J. (2009). *Globalization of an African Language: Truth or Fiction?* https://www.ajol.info/index.php/ljh/article/view/121549/111017 [perma.cc/P7UA-CVZ9]

Elswah. (2024). *Moderating Maghrebi Arabic Content on Social Media* (1). Center For Democracy And Technology. https://cdt.org/insights/moderating-maghrebi-arabic-content-on-social-media/ [perma.cc/3FGV-PTFV]

Githiora, C. (2002). Sheng: Peer language, Swahili dialect or emerging Creole? *Journal of African Cultural Studies,* 15(2), 159–181. https://doi.org/10.1080/1369681022000042637 [perma.cc/RB3L-A6RS]

Habwe, J. (2009). *The Role of Kiswahili in the Integration of The EastAfrican Region.* http://erepository.uonbi.ac.ke/handle/11295/28618 [perma.cc/HB6P-YWLM]

Kanijo, P. (2017). Code-Switching and Code-Mixing Errors among Swahili-English Bilinguals in Tanzania. *Kiswahili, 80(1*), Article 1. https://www.ajol.info/index.php/ksh/article/view/166001 [perma.cc/VGE5-2J8G]

Kannampilly, A., & Malalo, H. (2024, September 20). Kenya court finds Meta can be sued over moderator layoffs. *Reuters.* https://www.reuters.com/world/africa/kenya-court-rules-meta-can-be-sued-over-layoffs-by-contractor-2024-09-20/ [perma.cc/V424-S7K2]

Madung, O. (2022a). *Exporting Disinformation: How Foreign Groups Peddle Influence in Kenya through Twitter.* Mozilla Foundation. https://foundation.mozilla.org/en/campaigns/exporting-disinformation-how-foreign-groups-peddle-influence-in-kenya-through-twitter/ [perma.cc/J276-PB49]

Madung, O. (2022b). From Dance App to Political Mercenary: *How disinformation on TikTok gaslights political tensions in Kenya.* Mozilla Foundation. foundation.mozilla.org/en/campaigns/kenya-tiktok/ [https://perma.cc/6MAZ-7E6F]

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. https://www.taylorfrancis.com/chapters/edit/10.4324/9781003388289-4/silicon-savannah-digitising-marginalisation-job-mwaura [perma.cc/4L59-AV8U]

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/7EPT-V3GH]

Perrigo, B. (2022, February 14). *Inside Facebook's African Sweatshop.* TIME. https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/ [perma.cc/Q4F6-2DDG]

Perrigo, B. (2023, May 1). *150 AI Workers Vote to Unionize at Nairobi Meeting.* TIME. https://time.com/6275995/chatgpt-facebook-african-workers-union/ [perma.cc/L6BY-3RKC]

Roberts, S. T. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media.* Yale University Press. https://doi.org/10.2307/j.ctvhrcz0v [perma.cc/5R5A-ZQPF]

Shikali, C. S., & Mokhosi, R. (2020). Enhancing African low-resource languages: Swahili data for language modelling. *Data in Brief, 31,* 105951. https://doi.org/10.1016/j.dib.2020.105951 [perma.cc/2U75-3M3T]

Shiundu, A. (2023). Content moderation and countering disinformation in Africa- The tough choices. In *Taming the digital realm.* Friedrich Naumann Foundation. https://www.freiheit.org/sites/default/files/2023-08/taming-the-digital-realm.pdf [perma.cc/XL3U-R6HY]

Topan, F. (2008). *Language and National Identity in Africa* (A. Simpson, Ed.). Oxford University Press. https://academic.oup.com/book/48461/chapter-abstract/421398994?redirectedFrom=fulltext [perma.cc/G8JQ-UQGX]

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/YZ3X-TEX6]

Wanjawa, B., & Muchemi, L. (2021). Using Semantic Networks for Question Answering—Case of Low-Resource Languages Such as Swahili. In T. Ahram (Ed.), *Advances in Artificial Intelligence, Software and Systems Engineering* (pp. 278–285). Springer International Publishing. https://doi.org/10.1007/978-3-030-51328-3_39 [perma.cc/2822-Q4AA]

Wanjawa, B. W., Wanzare, L. D. A., Indede, F., Mconyango, O., Muchemi, L., & Ombui, E. (2023). KenSwQuAD—A Question Answering Dataset for Swahili Low-resource Language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.,* 22(4), 113:1-113:20. https://dl.acm.org/doi/10.1145/3578553 [perma.cc/4QX3-U878]

cdt.org

cdt.org/contact

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

202-637-9800

@CenDemTech

CENTER FOR
DEMOCRACY
& TECHNOLOGY