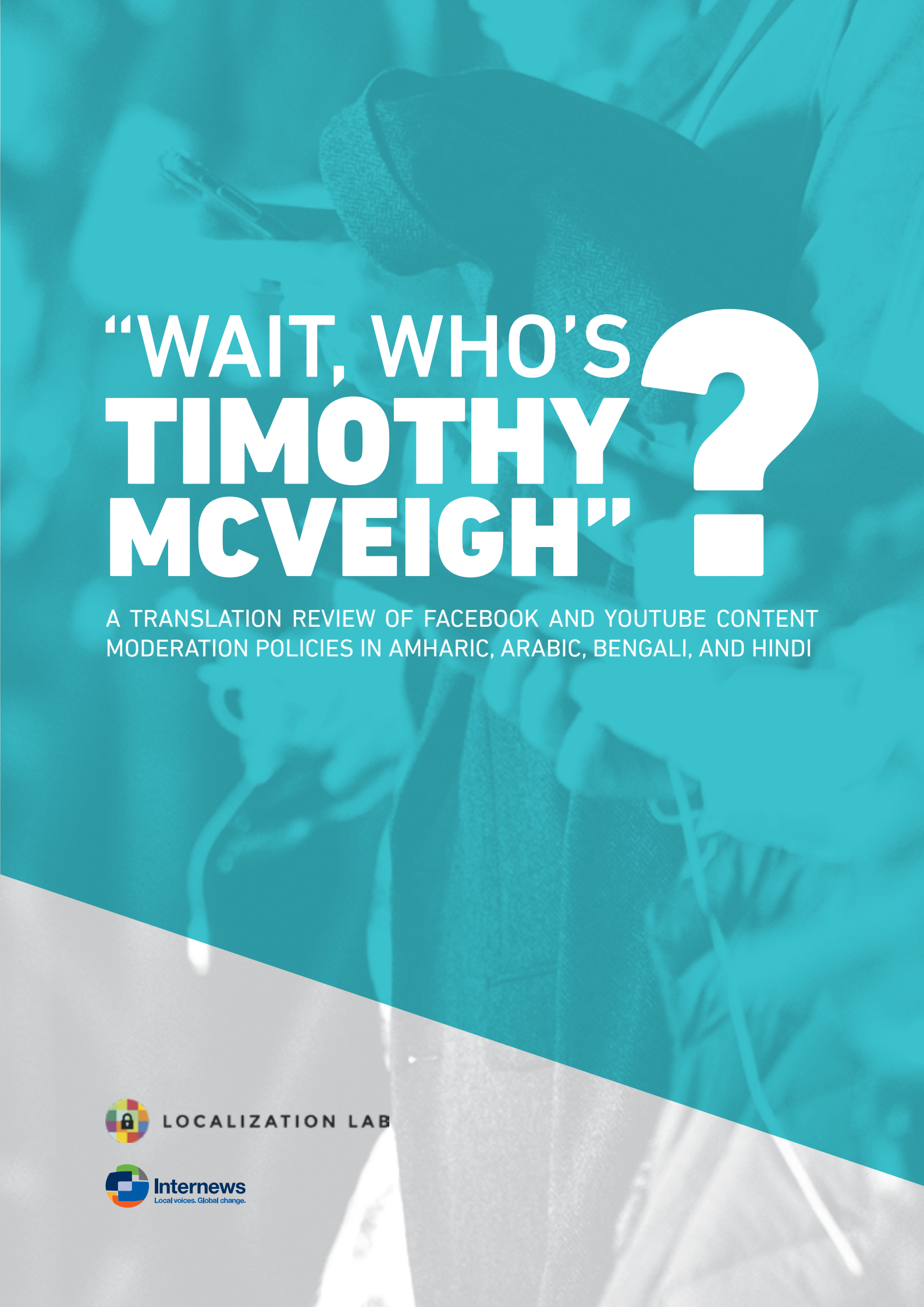


"WAIT, WHO'S TIMOTHY MCVEIGH"?

A TRANSLATION REVIEW OF FACEBOOK AND YOUTUBE CONTENT MODERATION POLICIES IN AMHARIC, ARABIC, BENGALI, AND HINDI

LOCALIZATION LAB

Internews
Local voices. Global change.

Contents

Figures

Tables

EXECUTIVE SUMMARY

This study has found that the translations of public facing content moderation policies in Amharic, Arabic, Bengali, and Hindi by Facebook and YouTube are far below a standard that would be considered acceptable by the average user. Each of the translations showed numerous and systematic errors in quality and usability that frequently impacted readers' ability to understand the policies without referencing the policy in the original source language. This effectively required readers to have advanced proficiency in English to understand the policy. Notably, the quality of the Amharic translations of Facebook's Community Standards was so poor that the translation reviewers relied on the English source language to comprehend the policy. The quality of Facebook and YouTube's translations in Bengali was of similarly poor quality and limited usability.

In conducting the reviews, the translators themselves had frequent questions about the references used to explain key policy concepts and terms. For instance, more than one reviewer asked, "Who is Timothy McVeigh?" when reviewing Facebook's content moderation policies. Timothy McVeigh was an American who committed the worst act of domestic terrorism to date during the 1995 Oklahoma City Bombing. His infamy has drawn support and followers among American hate groups, which social media platforms denounce. The problem, however, with references that lack localization and cultural adaptation to the target audiences is that it impedes readers' ability to engage with the translated policies because they are left wondering what the reference is.

Such impediments to the interpretability of the texts go against the basic premise of Facebook's Community Standards and YouTube's Community Guidelines. The policies are intended to inform users about what is and isn't acceptable on these platforms. Confusing and potentially misleading translations mean that it is impossible for non-English speaking platform users to make informed decisions about the content that they share or see on Facebook and YouTube. The resulting lack of user agency impacts all elements of content moderation and platform governance. Providing clear and usable translation of platform policies is a critical and achievable task that should be considered as a bare minimum requirement for companies with a significant international user base.

The problems in the translations highlight the need for Facebook and YouTube to communicate effectively with users beyond those in Anglophone countries. Neither Facebook nor YouTube provides public data on the number of end-users in each language that the platform supports, nor do they provide easily accessible data on users by country. This makes it difficult to estimate the impact of the current status quo on end users around the world. However, the current figures indicate that most end-users on Facebook and YouTube likely live outside of Anglophone countries.

The fact that such a significant percentage of platform users live in non-Anglophone countries underscores the need for quality translations. Compounding the poor quality of existing translations, there are dozens of languages in which content policies are not available in any form: Facebook supports 112 languages and provides policy translations in just 76, while YouTube supports 71 languages and provides policy translations in only 52. It is estimated that this impacts hundreds of millions of users, including in contexts with particularly high risks of harm. High quality translation of public facing policies should be an automatic step whenever any product is localized in a new language.

Key Findings

1. **The texts showed regular, often systematic, mistranslation of terms and translations that are not recognizable to speakers of those languages, even when the translated terms may be technically accurate.**
   As an example of such technically accurate, but semantically incoherent, translation, the Arabic translation of Facebook's incitement of violence policy includes a mistranslation of "call", a key policy term: the English-language original states that "calls", i.e., summons, to invoke violence are prohibited, while the Arabic translation refers to these "calls" as "phone calls". Consequently, the translated policy erroneously states that telephone calls invoking violence are prohibited rather than invocations to violence in general. A context-based translation would have recognized this semantic distinction, indeed this metaphorical use of the term "calls" in English, and it would have identified an adequate term or phrase to express this key term in Arabic.

2. **The lack of localized examples leads to confusion and insensitivity to audiences outside of the United States, to the exclusion of speakers from other Anglophone countries.**
   Key examples that explain and provide examples of prohibited content need to be localized for users in each language community, otherwise it can contribute to confusion and reflect a bias toward American culture. For example, Facebook's policy on dangerous individuals explicitly states that support and praise for dangerous individuals is prohibited and uses the example of the American domestic terrorist Timothy McVeigh, who is familiar to American end-users familiar with the 1995 Oklahoma City bombing. Consequently, a recurrent question throughout the policy reviews was "Who is Timothy McVeigh?" because this example was not resonant to end-users outside of the United States.

3. **Facebook's Amharic translation was assessed as unusable for most Amharic speakers. YouTube has removed their Amharic translation before an assessment was possible, possibly due to recognition of quality issues. Translations for other Ethiopian languages such as Oromo and Tigrinya are currently unavailable.**
   Facebook's Amharic translation showed systematic inaccuracies at the levels of word choice, grammar, and punctuation. Key policy terms and colloquial vocabulary were regularly mistranslated, which contributed to nonsensical and occasionally offensive statements in Amharic that did not carry the same meaning in the English original. Facebook's policy on adult nudity and sexual activity, for instance, states that "squeezing of female breasts" is allowed in the context of breastfeeding, while the Amharic translation translates this term as "crushing of female breasts" ("የሴቶችን ጡቶች መጨፍለቅ") which is inaccurate and potentially offensive.

4. **The review found that while Facebook and YouTube's Arabic language translations are mostly readable, they have numerous errors in the translation of key terms and the contextualization of the policies.**
   While the Arabic translations were mostly readable, they contained regular mistranslations that inhibited readers' ability to understand the policy. For instance, the translation of the terms discussing what is legal and lawful regularly used terms like "شرعي" ("shareiun") that carry a religious connotation. Instead, secularized legal terms like "قانوني" ("qanuniun") or "مشروع" ("mashrue") would be more appropriate and would reduce the risk of confusing people about whether the policy is referencing religious or statutory law.

5. **The review found that both Facebook and YouTube's Bengali policy translations are of limited quality and usability to most Bengali speakers because they do not reflect spoken or written language and they omit key terms.**
   The translation uses language that is abstract, passive and often opaque, which does not reflect the

spoken or written language of most end-users, and which is difficult to follow. The translations were often incomplete, with content like policy names, numbers, and key terms simply omitted. Furthermore, the Facebook translation regularly uses terms preferred by Hindu speakers of Bengali, but not by Muslim speakers of Bengali. This lack of standardization and use of terms unfamiliar to end-users across dialects can contribute to ambiguity about the meaning of the policy and lead readers to think that some groups are favored.

6. **The review found that Facebook's Hindi policy translations are of limited quality and usability to most Hindi speakers, while YouTube's policies were readable and mostly coherent, although they would be strengthened by using colloquial rather than "Hinglish" terms?**
   Facebook's Community Standards are of limited quality and usability to most Hindi speakers, due primarily to the inaccuracies in vocabulary and grammar and the reliance on English transliteration. The YouTube translations in Hindi were, conversely, stronger and employed generally short, simple, and clear sentences, although they also relied significantly on English transliteration to convey technical terms in Hindi.

7. **In languages that have significant regional variation, the translations used language that is dialectically specific and not universally recognizable, which limits the accessibility of the translation.**
   The risk of limited or no standardization is that minority groups can be marginalized. Lack of standardization may also indicate that there is a social bias toward a particular nationality, ethnicity, caste, class, or religion. For example, the Facebook translations used Bengali from (Indian) Central Standard Bengal as opposed to the dialect used by most Bengali speakers in Bangladesh, which suggests that there may be a bias toward localizing content for Bengali speakers in India rather than in Bangladesh, Relatedly, the Hindi translations on both Facebook and YouTube had significant "Hinglish" elements, which makes the text difficult to read for those who do not speak or read English, and which thus may reflect a social bias toward English-speaking Hindi speakers.

8. **The translations appeared to rely on machine translation which may be a factor in their limited readability. The use of machine translation without adequate human review is not acceptable practice.**
   There were numerous inaccuracies in vocabulary, grammar, and punctuation across the translations that were so regular as to suggest limited human involvement in the translation and review processes. The Amharic translation of Facebook Community Standards, for instance, showed consistent errors in translating singular and plural correctly, resulting in texts that do not reflect the correct numerical form. This error could have been prevented, or at the very least corrected, prior to the publication of the translations if there were more human oversight.

9. **The translations showed numerous errors of cultural and social sensitivity, such as gendered language, slurs, and offensive mistranslations, which were more frequent in the Facebook translations.**
   Taken as a whole, Facebook's translations had more errors of cultural sensitivity than YouTube's translations. Such issues of cultural insensitivity carry the risk of harm for end-users because it can contribute to statements contrary to those in the original policy and can lead readers to interpret that the policies condone harmful content. For example, the Amharic translation of Facebook's Community Guidelines on hate speech, for instance, used a term that many in Ethiopia consider to be a dog whistle for xenophobia and, thus, highly offensive.

10. **The translations frequently omitted key terms and entire phrases that resulted in incomplete translations, an issue that was particularly noticeable across YouTube's translated policies.**
    The omission of key terms and phrases, such as policy names, numbers, and various words throughout the policies makes the texts difficult to read and often uninterpretable for end-users. This issue of

publishing incomplete translations is avoidable if there is oversight in the translation and review process by an expert human speaker of the language.

11. **The translated policies regularly use terms, such as acronyms, without sufficient contextualization and that are recognizable often exclusively to an American audience.**
The lack of contextualization of key concepts into culturally relevant and identifiable terms is one of the fundamental barriers of interpretability for readers. For instance, Facebook's policy on dangerous individuals and organizations cites United States legal acronyms like "FTO", standing for "Foreign Terrorist Organization", that are incoherent as acronyms in the target language and confusing when left as acronyms in the Latin alphabet.

12. **The issues of cultural sensitivity and adaptation are preventable if platforms co-translate the texts with end-user communities and conduct reviews of the translations with end users' communities. The fundamental issue of translation usability is the contextualization of key concepts into culturally relevant and identifiable terms.**
The issues identified in this review are both preventable and rectifiable if platforms commit to including end-users in the translation and review of the policies. This is important because co-translation enables platforms to translate the policies efficiently and accurately and to adapt the policies so that they are easily understandable in the target language. This is a critical need because the current quality of the reviewed translations makes it frequently impossible for non-English speakers to engaged meaningfully with the policy content and to make informed decisions about what they can and cannot share on the platforms.

Overview of the Report

The following sections of the report document and discuss the key findings and primary issues of quality and usability in the translation of Facebook's Community Standards and YouTube's Community Guidelines. The sections aim to outline issues that are systematic throughout the translations and that create significant barriers to end-users' comprehension of the policies.

BACKGROUND

The advent of social media has democratized speech online by accelerating the dissemination of user-generated content. This has catalyzed the exponential growth and global ubiquity of social media platforms like Facebook and YouTube, both of which have over two billion users[1]. At the same time as user-generated content has contributed to this growth, it has also created significant liabilities when users share content that is misleading or false, promoting hate, celebrating violence, or facilitating exploitation[2]. As the volume of content increases, so too does the spread of harmful content, and technology companies have come under pressure by governments and the public to remove such content—and to do so quickly and at scale[3].

In response to this, technology companies have created content moderation policies that are often called "community standards" or "community guidelines"[4]. These policies outline the types of content that are prohibited on the platform. Yet the development of such rules to moderate the content of billions of users is a fundamentally challenging task. For one, the sheer scale of content to moderate means that governing content is reactive to existing content, rather than proactive in preventing future harmful content. For another, the "standard" of what is or is not acceptable is a subjective decision and, consequently, it is prone to implicit bias about what content is harmful and how it should be moderated[5].

Yet more fundamental than either of these challenges is the task of communicating the content moderation policies to the platform's entire user base in all users' languages. To date, there is no industry standard about how to translate public-facing policy documents, and different companies have varying translation priorities and quality thresholds. The result is that the quality of the translation may differ significantly depending on the objective of the policy and its translation: a translation aimed at communicating to an audience of end-users would look differently than one aimed at an audience of attorneys assessing platforms' liability for hosting harmful content.

This issue of a lack of industry standard in translation opens the key questions of this report. This report reviews the quality and usability of Facebook Community Standards and YouTube Community Guidelines in Amharic, Arabic, Bengali, and Hindi, and it considers how the translations impact user experiences. These platforms were selected because they are the world's two largest social media platforms, with approximately 2.96 billion monWthly active users on Facebook as of September 2022[6] and 2.6 billion active monthly users on YouTube as of July 2022[7].

While neither Facebook nor YouTube make end-user figures readily available, our estimates suggest that as many as 90% of Facebook[8] and 85% of YouTube users may live outside of Anglophone countries[9]. Moreover, India and the Middle East and North Africa (MENA) region have among the largest user bases for the platforms. In India, there are some 350 million monthly active users on Facebook and some 467 million monthly active users on YouTube, which makes it the single largest market. In Egypt alone, there are at least 45.9 million monthly active users on the platforms. This puts India and Egypt, the country with the largest population in the MENA region, among the regions with the largest reach and end-user base for either platform. Consequently, social media companies have a critical need to communicate effectively with the end-users outside of Anglophone countries and with end-users in these language communities specifically.

The platforms' content moderation policies have been translated into 76 and 52 languages respectively (see Figure 1)[10]. This leaves a conspicuous deficit in translated resources in the over 100 languages that the platforms support. The result is that is that the standards and guidelines are unavailable in languages with millions and even tens of millions of speakers. This deficit creates significant barriers to digital inclusion for people whose language is not translated on the site: the corollary is that this lack of adequate translation not only can keep people offline but also can inhibit equitable access to information about terms of use on the site.

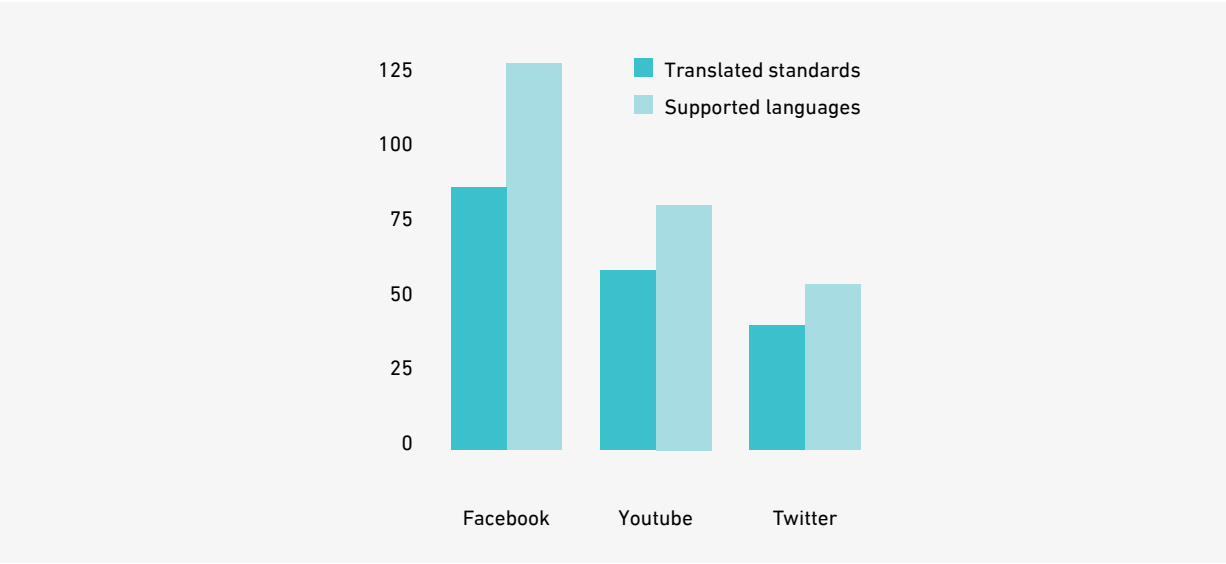


Figure 1 Supported languages vs. translated policies on top 3 social media platforms

This review is important for several reasons. The primary reason is that there are virtually no publicly available reports on the quality and usability of translated community standards and guidelines. This deficit exists despite robust documentation of operational issues in content moderation, which include limited resource allotment to content moderation in languages other than English[11] and serious human rights implications of harmful content that is barely moderated and sometimes not moderated at all. The Washington Post's Facebook Files, published in September 2021, show that Facebook has internally documented and tracked real-world harms that the content on its platform has exacerbated, while simultaneously ignoring warnings from employees about the risks to vulnerable communities that are exposed to harmful content[12]. Moreover, in September 2021, the Republic of The Gambia sued Facebook for failing to disclose materials relating to the incitement of ethnic hatred against Muslim-minority Rohingyas in Myanmar[13]. This indicates that Facebook is not only aware that poor content moderation carries potentially fatal risks to end-users and can implicate the platform in cases of crimes against humanity, but also that they care reticent to share their data or to make changes to their processes, even to support the discovery of other countries' crimes against humanity. Furthermore, it is noteworthy that YouTube has not faced the same legal pressure and allegations in relation to its content[14]. There is a lack of research about the content trends and patterns on the platform, in part because videos are more difficult to analyze en masse and in part because the platform provides few tools to do so.

Operationally, the issues of poor content moderation include limited human oversight and automated programs that inaccurately flag innocuous content and that systematically ignore harmful content[15]. The platforms regularly edit, change, remove and re-upload policies, as well as create new versions of them, and there is irregular archiving of these policies, particularly in translation. To date, YouTube provides no data on previous policies and updates to the content of its community guidelines in English or in any other language. While Facebook does provide this data, the archive of previous policies in each language that this study reviewed included policy content in English only.

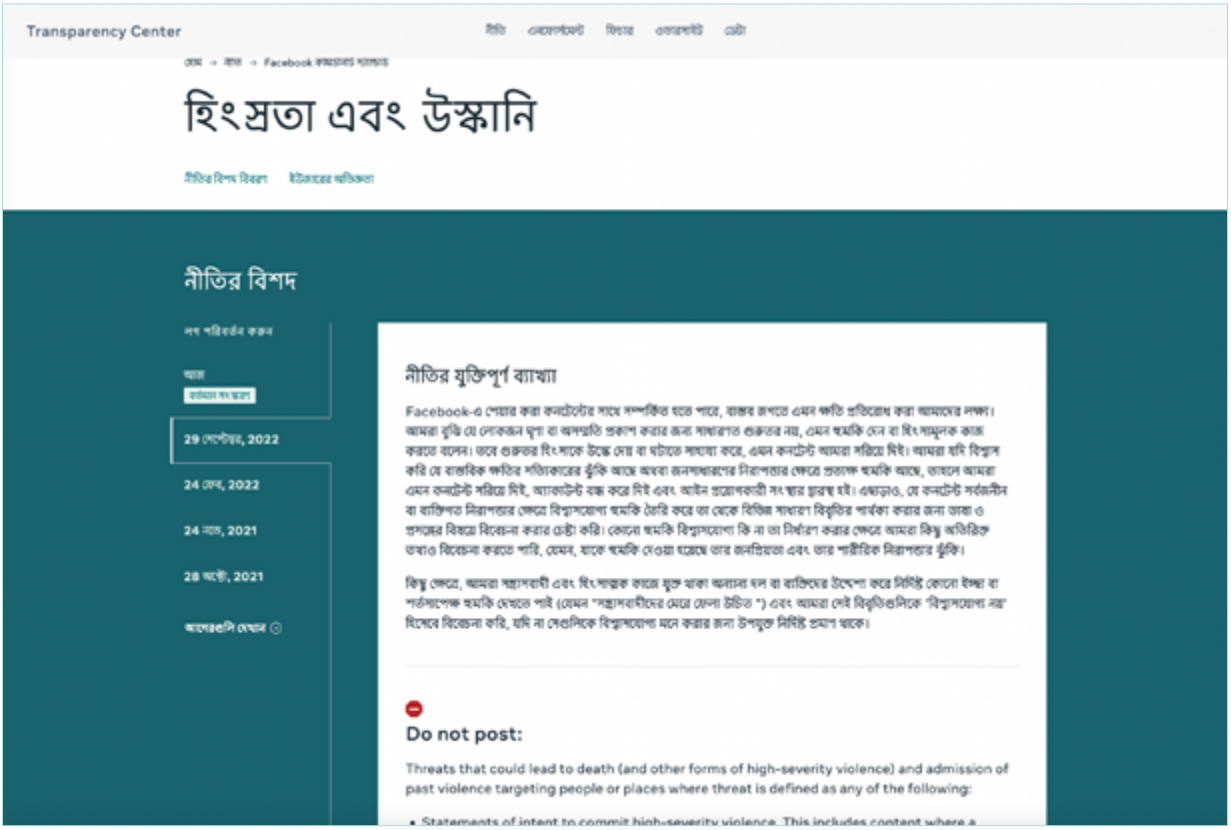


Figure 2 Archived Bengali-language Facebook Policy on Violence and Incitement from September 29, 2022

Both Facebook and YouTube rely on end-users to report harmful content, which creates a problem when the content moderation policies are poorly understood or do not exist in the user's language. Facebook's Help Center directs users to "report content" directly by clicking a "report" click in the original post and sending a short narrative description to Facebook about why the content is objectionable[16]. Likewise, the Google support page states that they "rely on YouTube community to report or flag content that they find inappropriate"[17]

For these reasons, understanding the quality and usability of content moderation policies translated from the English source language into target languages is a critical aspect of, indeed a prerequisite to, understanding the gaps and risks in content moderation. It is, in short, the first and most basic step to ensuring that there are commonly accepted minimum standards for communicating with billions of users online and improving content moderation and platform governance.

This Project

To explore the quality and usability of the translation of content moderation policies, Internews and Localization Lab reviewed the translations of Facebook and YouTube community standards and guidelines in four languages, Amharic, Arabic, Bengali, and Hindi, across 46 content policies spanning the two platforms. The study included reviews of all the policies in all languages, except for YouTube's Amharic translation, which YouTube removed shortly before the study's data collection phase.

The goal in this review is to develop an understanding of the status of policy translations in the industry and to evaluate their quality, usability, clarity, and accessibility. This document will serve as a baseline to anchor discussions with stakeholders and users about how community standards can be better communicated and about how the industry can develop commonly accepted minimum standards for the translation of public-facing content.

OBJECTIVE, SCOPE, AND METHODS

Objective

The objective of this review is to document the status of the translations of Facebook's Community Standards and YouTube's Community Guidelines, which are their official policies outlining the content that is prohibited on the site. The review evaluates the quality and usability of the translations in four languages. This document will serve as both:

- A proof of concept that establishes criteria for evaluating future translations
- An evidence base to anchor discussions with stakeholders and users

This document will serve as a baseline to inform conversations with stakeholders including developers, policy writers, and end-users about developing commonly accepted minimum standards in translation across the industry.

Scope

Internews and Localization Lab reviewed the Amharic, Arabic, Bengali, and Hindi translations of Facebook's 24 Community Standards and YouTube's 22 Community Guidelines to assess the quality and usability of the policies in each language.

Why We Chose These Languages

- We selected Amharic and Bengali because we had heard anecdotally that the translations in these languages were of very poor quality and that Bengali translation did not use the correct dialect for Bangladeshis. We wanted to assess their quality and usability issues.

- We selected Hindi and Arabic because India, where Hindi is most widely spoken, has the single largest user base on either Facebook[18] or YouTube[19] and Arabic, which is most widely spoken in the MENA region, has one of the largest per capita user bases across both platforms.

- Finally, we selected these languages because it is well established that social media content is contributing to social division, polarization, and humanitarian crises globally and that there are well-documented operational problems with the content moderation in these languages.

Table 1 Facebook Community Standards Reviewed

Facebook Community Standards Section	Policy
Violence and Criminal Behavior	Violence and Incitement
	Dangerous Individuals and Organizations
	Coordinating Harm and Publicizing Crime
	Restricted Goods and Services
	Fraud and Deception
Safety	Suicide and Self-Injury
	Child Sexual Exploitation, Abuse, and Nudity
	Adult Sexual Exploitation
	Bullying and Harassment
	Human Exploitation
	Privacy Violations
Objectionable Content	Hate Speech
	Violent and Graphic Content
	Adult Nudity and Sexual Activity
	Sexual Solicitation
Integrity and Authenticity	Account Integrity and Authentic Speech
	Spam
	Cybersecurity
	Inauthentic Behavior
	Misinformation
	Memorialization
Respecting Intellectual Property	Intellectual Property
Content-Related Requests and Decisions	User Requests
	Additional Protections of Minors

Table 2 YouTube Community Guidelines Reviewed

YouTube Community Guideline Section	Policy
Spam and deceptive practices	Fake engagement
	Impersonation
	External Links
	Spam, deceptive practices, and scams
	Playlists
	Additional Policies
Sensitive Content	Child Safety
	Thumbnails
	Nudity and sexual content
	Suicide and Self-Harm
	Vulgar language
Violent or Dangerous Content	Harassment and cyberbullying
	Harmful or dangerous content
	Hate speech
	Violent criminal organizations
	Violent or graphic content
Regulated Goods	Firearms
	Sale of illegal or regulated goods or services
Misinformation	Misinformation
	Elections misinformation
	Covid-19 medical misinformation
	Vaccine misinformation

Methods

Internews and Localization Lab, in partnership with two translation experts in each target language, developed criteria to assess the quality and usability of Facebook and YouTube content moderation policies in Amharic, Arabic, Bengali and Hindi. The criteria included eight categories, which are listed and defined in Table 3 below. This set of criteria formed the basis of the review rubrics, against which the translators assessed each translated policy.

The rubric included both quantitative and a qualitative component. The quantitative component consisted of a Likert scale (see Table 4), which is a 5-point psychometric scale used to measure a respondent's attitude, perception, and opinion through their level of agreement with a statement. Each number corresponds to a level of agreement: 1 corresponds to "strongly disagree" (with the statement); 2 to "disagree"; 3 to "neutral'; 4 to "agree"; 5 to "strongly agree". In this study, the respondents were asked questions about each of the criteria, such as "The meaning of the original policy in English is correctly carried over into the translation without omitting or changing the original meaning", to which they gave a response on the 1–5-point scale. The goal here was to gauge their overall assessment about how well the translation fulfilled each criterion.

The rubric's qualitative component consisted of a series of questions designed to gain contextual insight about reviewers' assessments, namely about how and why the translation did or did not meet the criteria in each category. In this section, the reviewers provided narrativized descriptions of their findings with concrete examples. Together, these documents informed the study findings, and the team identified the review findings using the observations detailed in these documents.

Study Limitations

While the study aimed to document the quality and usability of Facebook and YouTube's translated content moderation policies, the inductive nature of qualitative research means that individuals' subjective perspectives can impact what exactly they observe. The primary limitation of the study is that the review did not include the feedback of a representative sample of end-users. This type of feedback was out of the project's scope. For future research, however, developing an additional review component that includes working with a representative sample of users would be useful in understanding how users perceive the translation.

The translation-reviewers, as individual people, bring their own experiences, perspectives, and implicit biases to bear on their work, which can affect the universalizability of the findings. This is a component of all research—both qualitative and quantitative—and is important to consider while reading the report. This means that the study results are reflective of the quality and usability issues that the translators observed while using the rubric at a specific moment in time. It is not, therefore, an exhaustive list: there may, and indeed likely are, additional characteristics that impact end-user experiences, such as the level of readability, examples of prohibited content, and social descriptors that may become outdated, obsolete, or offensive. Consequently, the results of this study represent the observations of the team at a particular moment in time and may be considered a documentation of the most urgent translation needs of the document at the time of the report's writing.

Table 3 Criteria Used to Assess Translation Quality and Usability

Category	Definition
Accuracy	The meaning of the original policy in English is correctly carried over into the translation without omitting or changing the original meaning.
	Translation captures the meaning of the original policy in English without ambiguity.
Clarity of meaning	Translation is clear and can be understood by platform users, regardless of their digital literacy, and without creating any confusion.
Quality of expression	The translation reads naturally and is easy to follow for the platform users.
Consistency	The policy translation is consistent with translations of other policies on the platform, with regards to grammar, spelling, punctuation, tone, style, and gender.
	The regional language or dialect used in the policy translation is consistent with the translations of other policies of this platform.
	The inconsistency in the grammar, style, tone, and terminology in the translated policy may be the result of machine translation.
Inclusivity, Equality, and Diversity	The policy translation takes into consideration the diversity of the language community in race, ethnicity, culture, and religion and avoids associating the translation with a certain religion or ethnicity that originates from those countries. For example, providing phrases or examples about the Muslim community in the Persian translation.
	The policy translation is free from phrases or examples of sexism or assigning traditional gender roles. For example, translating the word parent in "parents are accountable for governing children's use of the platform", as mothers.
	The policy translation is sensitive and respectful to vulnerable and marginalized groups, such as indigenous communities, the seniors, LGTBQ+, victims of dangerous organizations, and survivors of suicide attempts.
	The policy translation is sensitive and respectful and does not discriminate against people with disabilities or uses terms that emphasize disabilities, e.g., the blind, the disabled.

Table 4 Likert Rating Scale used in Review

Rating Scale	Meaning	
1	Strongly Disagree	Statement is very inaccurate, translation is not acceptable and misleading
2	Disagree	Statement is at times inaccurate, the translation is of poor quality, and could create confusion
3	Neutral	Statement is accurate, the translation is acceptable but does not make sense
4	Agree	Statement is accurate, the translation is of good quality, and could be understood
5	Strongly Agree	Statement is very accurate, the translation is of excellent quality and could be understood easily

AMHARIC TRANSLATION REVIEW FINDINGS

Summary of Facebook Community Standards Review

Internews and Localization Lab's review of Facebook's Community Standards in Amharic showed that all 24 community standard policies were translated from English to Amharic and that the translated policies:

1. **Appeared to rely on machine translation, which produced texts that were inaccurate, frequently incoherent, and had systematic errors in grammar, spelling, and punctuation;**

2. **Inconsistently used Addis Ababa dialect, on which standard written Amharic is based, and the translation unpredictably used terms from other dialects;**

3. **Regularly used gendered language that contributes to gender-based stereotypes rather than gender-neutral language;**

4. **Frequently lacked contextualization of key policies, lacked affiliate links in Amharic, and included errors that create meanings at odds with the policy's intention.**

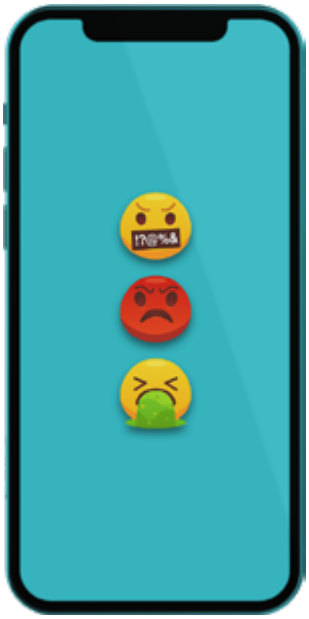
The review findings show that Facebook's Community Standards are of limited quality and usability to Amharic speakers due primarily to the systematic inaccuracies in vocabulary, grammar, and punctuation. This limited quality appeared to be caused by limited human oversight. This resulted in texts that were confusing and often incoherent. The quality was so poor that it prompted our translators to rely on the original English language policies to interpret the meaning of the texts in Amharic. This indicates that the Community Standards are unlikely to be usable for most native Amharic speakers, and even less so among speakers for whom Amharic is a second or third language.

The findings show that the Amharic translation was often unintelligible and lacked the contextualization and cultural references necessary to communicate the policies in clear Amharic prose. For instance, the title "Community Standards" is mistranslated as "የማህበረሰብ ደረጃዎች". ("yemahiberesebi derejawochi"), which translates roughly to "levels of society" rather than "community standards". Instead, a better translation would be "የማህበረሰብ ደንቦች" ("yemahiberesebi denibochi"), which translates more closely to "community rules". Here, the term "community" in English would be better rendered as "societal", aligning more closely with the way that Facebook users in Ethiopia discuss and conceptualize Facebook—not as a space that is proximal and part of their everyday lives, but as a space reflecting broader Ethiopian society. This is an important point because while the translation does literally the English language, the result is that the title does not reflect how Amharic speakers refer and understand Facebook's platform as an entity in their lives. This creates a lack of clarity to readers engaging with the policies.

The translations used word to word translations that resulted in systematic inaccuracies at the levels of word choice, grammar, and punctuation. Both key policy terms and vocabulary in the body text were regularly, and consistently, translated incorrectly and, occasionally, in ways that read as nonsensical and offensive. For example, in the policy on adult nudity and sexual activity, Facebook states that the "squeezing female breasts" is allowed only in the context of breastfeeding. The problem is that the phrase "squeezing female breasts" is translated as "የሴቶችን ጡቶች መጨፍለቅ", which states that "crushing female breasts" is prohibited. This mistranslation from "squeezing" to "crushing" in Amharic is both inaccurate and potentially offensive.

There are systematic errors in grammar and punctuation that contribute to incoherence in the Amharic text. For example, the translations make consistent errors in translating singular and plural forms correctly. This results in texts that frequently do not reflect the numerical form in the original and that lack clarity of meaning for readers in Amharic. Moreover, there were also systematic errors in punctuation, such as in the

use of the English comma (“,”), which is used to separate items on a list or to distinguish clauses lists, and the colon (“:”), which is used to precede a list. These types of punctuation are used, but the punctuation mark is distinct in Amharic. Instead, the translations should use the punctuation mark “፣”, which is used analogously to a comma in Amharic and the punctuation mark “:-“, which is used analogously to colons in English. This type of systematic error, while perhaps minor in comparison to lexical incoherence, does create difficulties and discomfort for the reader who must either know how English language punctuation works in English, must look it up otherwise, or who may resort to guessing its meaning, thus creating opportunities for misunderstanding, and unsuccessfully communicating the policies in correct Amharic.



The policies’ lack of social contextualization to Ethiopia also contributes to lack of coherence for readers in Amharic and can have the unfortunate consequence of communicating a meaning opposite to what the policy intends. The mistranslation of key terms in the policy of hate speech provides a case in point. The policy states that “We also protect refugees, migrants, immigrants, and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies.” The Amharic translation of the word “migrants” uses the term ““ሞጤ” (“met’ē”), which carries a meaning closer to “foreigner” and is considered highly pejorative by Amharic speakers. Instead, a better translation would use a term like “ፍልሰተኞች” (“filisetenyoche”), which translates more closely to “immigrants” and is considered respectful. The term “ሞጤ” is widely used in the country’s civil conflict by politicians and political elites who exploit people’s economic grievances and who urge them to take up arms to protect their land. This mistranslation thus has the unfortunate and ironical result of communicating that certain forms of derogatory speech may be permissible on the site, even when this is contrary to what the policy states in English.

Table 5 Summary of Facebook Community Standards Review of Amharic Translation

Criteria	Finding	Example
Accuracy	The translations used word for word machine translations that frequently resulted in semantically incoherent, if technically accurate, texts.	The title "Community Standards" is translated as a concatenation of the terms "community" and "standard, which diverges from the ways Amharic speakers describe the platform; rather a term aligning with "society" or "societal" that would resonate better with end-users.
Errors	The translation made systematic errors in grammar and punctuation.	The comma punctuation mark, used in English, is unfamiliar to Amharic speakers who use an analogous mark "፣ " to distinguish items on a list or separate clauses. Likewise, the translations routinely switch singular in English with plural in Amharic and vice versa.
Clarity of meaning	The meaning of key concepts in the original source language frequently diverges from the meaning rendered in the Amharic translation.	The translation of the phrase "squeezing female breasts" in English, in reference to depictions of breastfeeding a child (permitted on the platform) was rendered as "የሴቶችን ጡቶች መጨፍለቅ" ("yesētochini t'utochi mech'efilek'i"), literally "crushing female breasts", in Amharic.
Quality of expression	The sentence structure in Amharic frequently lacks syntax or phrasing that resembles natural language.	The translation of "personal contact information" as "የግል የእውቂያ መረጃ" ("yegili ye'iwik'Īya mereja") translates each English term separately and creates a nonsensical concatenation of the three terms. A better translation would refer to this phrase as "የግል መገኛ አድራሻ" ("yegili megenya ādirasha"), or "personal address".
Consistency	The low accuracy, clarity and the high number of errors were consistent across all the translated policies, and the same terms were regularly translated in diverging ways,	The term "access", e.g., from the policy's statement about engaging in "unauthorized access" is translated inconsistently as "መድረሻ" ("mediresha") and as "መዳረሻ" ("medaresha"). "Access" is translated as "mediresha" roughly translating to "arrival" or "destination" and then as "medaresha", roughly translating to "range", when the second translation could have been used in both instances.
Diversity, equality, and inclusion	The translated policies sometimes used terms terms that Amharic speakers consider to be derogatory, inflammatory, and offensive.	The translation of the term "migrant" from English to Amharic uses "ሜጤ" ("met'ē"), more closely translating to "foreigner") that is considered highly derogatory.
User accessibility, register and tone	The translated policies are written at a higher readability that creates barriers to comprehension for readers with basic reading comprehension.	Much of the correctly translated policy uses technical or literary vocabulary that is unfamiliar to most readers.
Contextualization	The translation omitted cultural adaptation, and affiliate links and resources were rarely translated into Amharic.	The translated policy makes reference to hate groups, e.g. the "KKK", which is recognized as the hate group par excellence to Americans, but whose acronym and organization are not recognized by a wide Amharic speaking audience.

ARABIC TRANSLATION REVIEW FINDINGS

Summary of Facebook Community Standards Review

Internews and Localization Lab's review of Facebook's Community Standards in Arabic showed that all 24 community standard policies were translated from English to Arabic and that the translated policies:

1. **Overall produced texts that were readable, although they contained errors in spelling, grammar, and punctuation;**
2. **Routinely used literal translation and employed words that were unfamiliar and highly technical;**
3. **Contextualized key policy concepts in terms most familiar to readers in Anglophone countries omitting explanations in terms familiar to readers in the MENA region.**

The review findings indicate that Facebook's Community Standards were generally readable in Modern Standard Arabic (MSA) and occasionally had incorrect use of grammar, punctuation, and spelling that detracted from readability. While the text was readable to users familiar with MSA, there were errors in the copy that lowered the quality of the translation in Arabic. These include the forward slash (i.e., "/", instead of "أو", which is used analogously to the forward slash in English. Such errors can make the texts more difficult to read and reduce the quality of the translation.

One of the main issues with the Arabic translation was the frequent mistranslations of key terms that created ambiguity in the meaning of the text. For example, in the policy on violence and incitement, there were numerous inaccuracies in the translation that communicate a meaning other than what the English-source text implied. For example:

- In the policy on violence and incitement, one of the clauses mistranslated the word "calls" about calls to action, as "phone" calls. The policy states in English that "calls to action" that promote violence are prohibited and that "We may also restrict calls to bring armaments to certain locations". In Arabic, the term "calls" was mistranslated as " المكالمات" ("almukalamat"), which is better translated as "telephone" calls. This creates a serious inaccuracy. First, it communicates that only "phone calls", rather than summons to violence more generally, will be restricted, thus implying that the latter may be acceptable. Second, in mistranslating calls as phone calls, the policy statement implicitly states that users' Facebook audio calls on the platform may be monitored, which conflicts with their other policies stating that calls can be end-to-end encrypted.

- Similarly, the translation of the terms discussing what is legal and lawful regularly used terms in Arabic that carry religious connotations, such as "شـرعي" ("shareiun"). Secularized legal terms like "قانونـي" ("qanuniun") or "مشـروع" ("mashrue") would be more appropriate and reduce the risk of confusing people about whether the policy is referencing religious or statutory law. Likewise, the translation of terms like "law enforcement officers" was ambiguous. In this case, the translation referred to them in Arabic as " سلطات إنفاذ القانون" ("sulutat 'iinfadh alqanun") which translates more closely to "law enforcement authorities" and creates ambiguity about who exactly has the authority to enforce the law and whether this includes ordinary citizens.

The policies' lack of contextualization and the platforms' history of poor content moderation in the MENA region open questions about the resources available to moderate content. For instance, the policy on violence and incitement states that posting violent content is prohibited with some exceptions:

> Statements admitting to committing mid-severity violence except when shared in a context of redemption, self-defene (sic), fight-sports context or when committed by law enforcement, military or state security personnel.

Here, the original text and the translation leave critical terms open for users to interpret and define. For instance, the meaning of acceptable "mid" severity violence is ambiguous because it does not define what exactly is "mid" in comparison to "low" or "high" severity violence. Likewise, the translation of "redemption" is confusing because the translation in Arabic used a term more closely translated to "salvation". The consequence of this mistranslation is that readers in Arabic would have reason to think that the document is saying that violent content committed in an act of "salvation" is permissible.

The policies frequently explain key terms or use examples of prohibited content that resonate with end-users in Anglophone countries. For example, the policy on dangerous individuals and organizations states that content praising or in support of dangerous individuals like "Timothy McVeigh" will be removed from the platform. The issue here is that many end-users outside of the United States may be unfamiliar with Timothy McVeigh, who was an American domestic terrorist responsible for the 1995 Oklahoma City bombing, which is, to date, the deadliest act of domestic terrorism in the United States[20]. While this contextualization may resonate with end-users from the United States, it is much less likely to resonate to end-users outside of the context of the United States.

Moreover, the policies' lack of contextualization for readers in the MENA region means that there are systematic ambiguities that highlight the inequitable enforcement of content moderation policies. Many Palestinian users of Facebook, for example, are living under a military occupation and are often targets of un- or under- moderated hate speech circulated across the platform[21]. This opens the question about the mechanisms that exist to moderate hate speech and depictions of violence, such as against people living in extreme social, political, and economic marginalization and against people living in contexts where state security organizations perpetuate human rights abuses against ordinary citizens.

Table 6 Summary of Facebook Community Standards Review of Arabic Translation

Criteria	Finding	Example
Accuracy	Key terms throughout the text were frequently technically accurate, but incorrect for the context.	The translation of the term "calls", e.g., "calls" to violence, in the policy on incitement of violence, is translated as "المكالمات" ("almukalamat"), or "phone" calls, which creates confusion because the translated policy consequently indicates that some "phone" calls may be restricted.
Errors	The policies had occasional typos throughout the texts.	The terms "dangerous individuals" and "dangerous organizations" ("الأفرادالخطيرون والمنظمات الخطيرة") are misspelled in the policy on dangerous individuals and organizations.
Clarity of meaning	The policy and the translation do not always define exactly what eacWh term means, which can create ambiguity for end-users reading the policy.	The policy on incitement of violence delineates "high", "mid" and "lower" severity violence, without adding definitions or practical implications for what these three levels of violence involve.
Quality of expression	Sentences structures often reflect English-language syntax in Arabic, which contributes to ambiguity and a lack of natural expression in Arabic.	The sentence structure, particularly in the section on firearms prohibitions, reflects American legal English, which creates sentences that do not resemble Arabic syntax.
Consistency	Most inaccuracies were consistent throughout the policies, the regularly use of technically accurate, if semantically incoherent or ambiguous, terms and phrases.	The translation of the policy on incitement of violence prohibits organizing violence against individuals who hold a protected status (e.g., race, ethnicity, religion, sexuality, gender) the translation of which is unclear in Arabic. In particular, the translation describes protected individuals as having "attributes with protected rights", which carries an ambiguous meaning for readers in Arabic.
Diversity, equality, and inclusion	The translation used terms that can create the perception that some people are exempted or singled out by the policy.	The translation of the term "blasphemy" uses the term "كفر"" ("kafar"), which is a term specific to Islam rather than a term like "تجديف" ("tajdif"), which can be used in reference to any religion.
User accessibility, register and tone	The translation uses Modern Standard Arabic, the standard written form of Arabic, although the translation incorporates culturally specific acronyms in English to reference specific organizations.	The translation uses the acronyms "FTO" ("Foreign Terrorist Organization"), "SDN" ("Specially Designated Nations" and "SDGT" ("Specially Designated Global Terrorists"), which are US Government specific terms.
Context	Most affiliate content links lead to English-language resources and explain key concepts with references familiar to end-users in Anglophone countries	Links of the List of Tools on Facebook and Bullying Prevention Hub lead to an English-language source and use cultural references , such as a reference to the "Illuminati", that are most familiar to

Summary of YouTube Community Guidelines Review

Internews and Localization Lab's review of YouTube's Community Guidelines in Arabic showed that all 22 policies were translated from English to Arabic and that the translated policies:

1. **Overall were coherent and employed mostly correct grammar, punctuation, and spelling;**
2. **Regularly used terms that were technical and made reference to cultural and Internet phenomena most familiar to end-users in Anglophone countries, esp. the United States;**
3. **Inconsistently linked to affiliate resources in Arabic and lacked contextualization of resources in Arabic.**

The findings of the review of YouTube's Community Guidelines in Arabic indicate that the translations were generally coherent and were written in MSA. The translations employed mostly correct grammar and that had only a few errors in punctuation and spelling. However, the method of word for word translation of the source text from English to MSA meant that the translated text frequently had phrases or clauses that resemble English expression, rather than Arabic expression. This included lengthy sentences with multiple subordinate clauses that are difficult to follow in Arabic and require a high level of reading proficiency for readers in Arabic.

The translated policies can also make gendered distinctions and address readers using gendered terms that are not reflected in the original English source language. For example, statements where "you", which is gender-neutral in English, was the subject in English were often translated in gendered ways in Arabic, e.g., as "أنتم" ("antum"), meaning masculine plural "you". A gender neutral or gender inclusive choice that addresses all end-users and that does not use gender-based stereotypes would improve the quality of the translation.

The lack of clarity in the English source text also contributed to ambiguity in the translated Arabic texts. For example, the English language policy states that the "use of excessive profanity" or the "use of heavy profanity" are prohibited on the platform. This distinction creates ambiguity because it suggests that "moderate" profanity may be acceptable, and it does not define "excessive" and "heavy". Similarly, the policy on harmful or dangerous content policies prohibits instructions on how to prepare bombs and includes a list of types of bombs, e.g., pipe bombs, cigarette bombs, and Molotov cocktails. Such terms, however, are most familiar to readers in Anglophone countries, and their translation in Arabic is confusing. The term "pipe bomb", for instance, is translated as "tube" bombs; cigarette bombs are translated as "fumigation cigarettes", and Molotov cocktails may be unclear for many readers. The result is that the policy lacks clarity of meaning and contextualization for Arabic speakers on Facebook.

Table 7 Summary of YouTube Community Guidelines Review of Arabic Translation

Criteria	Finding	Example
Accuracy	The translation mostly used terms familiar to readers, however, key terms like "fake" and "false" sometimes were translated interchangeably.	The translation of the term "false claims" was translated as "ادعاءات زائفة" ("aidiea'at zayifatun"), which translates more closely to "false allegations". The translation needs to be careful to translate "false" and "fake" differently as "مغلوطة" ("maghluta") and "زائفة" ("zayifa").
Errors	The translation did not always use punctuation, in particular, when translating lengthy sentences from English into Arabic. While not an error, it can contribute to ambiguity and lower the readability of the text.	Sentences with more than ten words generally lack sufficient punctuation in Arabic.
Clarity of meaning	The translation of specialized vocabulary does not always include an explanation, which can contribute to confusion about what the text means.	The transliteration of terms like "hydroxychloroquine" as "هيدروكسي كلوروكوين" or "haydruksi kulurukuin" requires explanation, as this may not be a term familiar to most readers.
Quality of expression	The quality of expression is generally proficient and reflects Arabic syntax and formal ways of writing.	The use of Arabic syntax and standardized language means that many end-users can engage with the document.
Consistency	The translation is clear and there are few errors in the copy; where errors exist, they are not consistent or systematic.	The typos throughout the translated text are few and inconsistent.
Diversity, equality, and inclusion	The translation occasionally used terms that are gender exclusive by using a masculine gender that could be phrased in a gender neutral way.	The plural form of the pronouns "you" and "they" in Arabic distinguishes between masculine and feminine forms of the noun, specifically as "أنتم" ("antum") and "أنتن" ("antunna") being the masculine and feminine forms of plural you and "هم" ("homa") and "هن" ("hunna") being the masculine and feminine forms of "they" in Arabic.
User accessibility, register and tone	The translation used Modern Standard Arabic, which is widely accessible to readers in Arabic speaking countries, as the written language is standardized.	The translation of the policy does not adopt regionally specific terms across the Arabic dialect continuum, which increases the readability of the text.
Context	The links to affiliate information and resources regularly lead to resources in English.	The links to resources on accurate, up to date, vaccine information on WHO and UN websites led to the resources page in English.

BENGALI TRANSLATION REVIEW FINDINGS

Summary of Facebook Community Standards Review

Internews and Localization Lab's review of Facebook's Community Standards in Bengali showed that all 24 community standard policies were translated from English to Bengali and that the translated policies:

1. **Appeared to use machine translation and phonetic transliteration to convey key policy terms that created technically but not functionally accurate renderings of the original source text;**

2. **Primarily used Central Standard Bengali which is based on an Indian dialect of Bengali that users terms that are not common across the dialect continuum and that includes phrases preferred by Hindu but not Muslim speakers of the language;**

3. **Frequently explained prohibited content with references relevant to users in Anglophone countries and examples that can be illogical, confusing, and offensive to both Hindu and Muslim speakers of Bengali.**

The review findings indicate that Facebook's Bengali translations are of limited quality and usability to most Bengali speakers. This inhibits their ability to engage meaningfully with the content of the policies, such as knowing which types of speech are prohibited or allowed. In particular, the translations use language that is abstract, passive, and at times opaque and that reads as formal and academic. This does not reflect the spoken or written language of most end-users and is thus unnatural and difficult to follow. Further complicating this issue, the translations were frequently incomplete—transliterating or leaving content, including policy names, untranslated, which creates ambiguity for readers in Bengali.

> ! Machine translation is considered poor practice in translation because it inaccurately renders natural language and produces texts that can reflect discriminatory stereotypes. Anti-Muslim bias, for example, in language models is a persistent problem. In 2021, Abid et al. found that GPT-3, a state-of-the-art contextual language model, "consistently and creatively" analogized "Muslim" with violence in over 60% of test cases and specifically with "terrorist" in 23% of test cases[22]. These biases are severe even when compared to biases about other religious groups, such as anti-Jewish tropes mapping "Jewish" with "money" in 5% of test cases. Using the 6 most positive adjectives to overcome anti-Muslim bias with adversarial text prompts reduces violence analogies from 66% to 20%, which is still significantly higher than negative associations for other groups.

The findings also show that the use of Central Standard Bengali as the target language for the translation uses terms that are regionally specific to West Bengal (India) rather than across the Bengali dialect continuum in India and Bangladesh. The translation of the term "water", for example, is regionally specific and there are several distinct phonemes used to denote "water" in Bengali dialects. More concerning, these types of translation choices can reflect and contribute to social conflict, for which Facebook's translation of "religious sacrifice" provides a case in point. Both Hindu and Muslim speakers of Bengali translate the term differently and do not use the terms interchangeably. Facebook's selection of the Hindu-preferred term to denote "religious sacrifice", which are not allowed to be depicted on the platform, carries the risk of alienating Muslim users by contributing to confusion about exactly which religious sacrifices are prohibited and by contributing to the perception that certain people and dialects are preferred. While water may appear to be a pedestrian example for translation, it indicates that the translators need to consider the entire range

terms to communicate in Bengali, from prosaic, everyday terms like "water" to highly specific terms like "religious sacrifice" when speaking to the platform's audiences.

The examples of prohibited content in the policy translations require regional and cultural contextualization. Currently, the policies are contextualized for Anglophone audiences, which means that the examples used to demonstrate prohibited behavior are often irrelevant and can carry very different meanings to Bengali speakers. In particular:

- A significant portion of the policy on restricted goods and services discusses the use and promotion of firearms, which are highly regulated in Bangladesh and India and not a protected right as they are in the United States.

- The policy refers to social norms that are not widespread in Bangladesh or India. For instance, the standards state that sharing images of Bengali women without a veil constitutes "harm", however, wearing a veil is not mandatory for Bengali women.

- Examples of derogatory speech in the policy on Hate Speech refer to abusive speech familiar to English speakers, e.g., to terms like "cows", "monkeys", and "potato". These terms do not translate to a Bengali linguistic context, where cows, for example, are venerated by Hindus as a representative of divine and natural beneficence.

The consequence of this lack of contextualization is that the policies are ambiguous, confusing, and, at times, illogical to Bengali speakers.

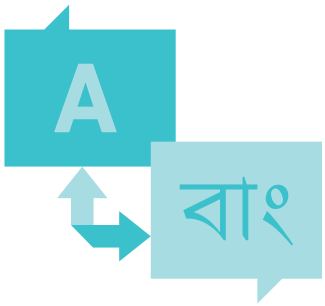
The review findings also show that the policies referred to religiously motivated violence by Muslims to exemplify terrorism and violence. The policy on violence and incitement provides the example that "Those fighting for the Islamic State are truly brave!" is an instance of prohibited incitement of violence. While such a statement could be used to incite violence, there is robust evidence that Muslim terrorism is significantly less prevalent than many Americans, Canadians, Australians, New Zealanders, and Europeans consider it to be[23]. The problem with this example is that it singles out Muslims and contributes to negative stereotypes about Muslims when this type of example can be described in concrete, yet generic terms.

Table 8 Summary of Facebook Community Standards Review of Bengali Translation

Criteria	Finding	Example
Accuracy	Machine translation appeared to be used systematically throughout the document, rendering technically accurate but functionally low quality translations.	The translation of "immediate family members" as "অবিলম্ব পরিবারের সদস্য" ("abilamba paribārēra sadasya") is closer to saying "quick" or "prompt" member of family and does not reflect social categories that Bengali speakers use.
Errors	Most of the translations made relatively few spelling errors, but used vocabulary that is uncommon and irregular.	The term "marijuana" was transliterated into Bengali as "মারিজুয়ানা" ("mārijuẏānā") when most Bengali speakers refer to marijuana as "গাঁজা" ("gāmǰā") or "ভাং" (bhāṁ").
Clarity of meaning	The inaccurately translated words contribute to ambiguous and confusing meaning of key policy terms.	The memorialization policy translates "victims of murder and suicide" as "আক্রান্ত" ("Ākrānta"), which more closely translates to "afflicted" or "infectecd", carrying a connotation of disease. A better translation would use "শিকার" ("Śikāra"), or "victim", indicating that the person was harmed as a result of a crime, accident or other event.
Quality of expression	The translation in Bengali used abstract, passive, and at times opaque language that reads as formal and academic and that did not reflect either the spoken or written language of most end users and is thus unnatural and difficult to follow.	The clause "public to friends-only" in the memorialization policy about how a designated person can change the privacy settings on the account of a deceased person is translated as "সবাই থেকে ফ্রেন্ড", meaning something more like "from friend to all". A better translation would state "সার্বজনীন থেকে বন্ধুদের জন্য শুধু", or "from public to friends only".
Consistency	There is inconsistent translation of key terms throughout the policies, where the same terms and concepts are translated differently throughout the document.	In the policy on suicide and self-injury, "self-harm" was translated as "স্ব-আঘাত" ("sba-āghāta"), "self-injury", and in some other places it is not translated into Bengali.
Diversity, equality, and inclusion	The language of the policies is written in Central Standard Bengali which most resembles language spoken in India and used terms preferred by Hindus.	The policy on incitement of violence cites "places of worship" as locations where violence is strictly prohibited. The Bengali translation refers to this as "উপাসনাস্থল" ("upāsanā sthala"), which is a Hindu place of worship and a term that Muslims, who worship in "মসজিদ" ("masajida"), i.e. mosques, do not use.
User accessibility, register and tone	The document is primarily translated into Central Standard Bengali, and there are few translations across the dialect continuum. The document is written in a formal tone creates a barrier to readers with varying levels of literacy.	The clause "which would typically require the individual to self report" from the policy on memorialization is translated into a sentence that reflects legal English in its use of passive voice, but that is difficult for most Bengali speakers to understand. Use of clear subjects, verbs, and objects in the English source language would limit opacity.
Context	The policies provide explanations using examples familiar to readers in North America and include affiliate links to civil society organizations serving US citizens.	The policy on child sexual exploitation and nudity refers to the National Center for Missing and Exploited Children, which is a US nonprofit organization serving the American public and which has no presence in Bangladesh or India.

Summary of YouTube Community Guidelines Review

Internews and Localization Lab's review of Facebook's Community Standards in Bengali showed that all 22 community standard policies were translated from English to Bengali and that the translated policies:



1. **Appeared to use machine translation to render text from the English source language into Bengali;**
2. **Regularly used terms that, while standard and recognizable to most Bengali speakers, were nonetheless ambiguous in meaning;**
3. **Inconsistently explained prohibited content with references relevant to Bengali speakers in Bangladesh and India and regularly linked to organizations serving the American public.**

The review findings show that YouTube's Community Guidelines are of limited quality and usability to speakers of Bengali. Despite the fact that most of the translated language is technically accurate, the translated language reads as unnatural and abstract to Bengali speakers. This lack of fluency in the translation inhibits readers' ability to interpret what the policy means. Moreover, the translations were frequently incomplete and systematically incorporated transliterated English written in the Bengali alphabet. Numerous policy names, clauses, terms, and even full sentences remained untranslated. This not only exhibits an incomplete translation, but it also creates significant barriers for Bengali language audiences to understand the text.

The findings show that the use of machine translation to render content from English into Bengali produced translations that are incomplete, erroneous, abstract, and that routinely omit key terms in the policies. For example, in the policy on impersonation, the translated policy omits numbers that existed in the English source language. Likewise, important terms like "আত্মহত্যা", meaning "suicide", and prepositions like "জন্য" were frequently misspelled, which creates a significant lack of clarity and professionalism in the translation.

The lack of contextualization of key policies is noteworthy because it may contribute to readers' perception that the policy does not apply to them, and this can contribute to confusion about the intended policy audience. The firearms policy, for instance, discusses depictions of firearms in ways that are specific to the United States' epidemic of gun violence. In particular, the policy prohibits an Internet-based stunt called the "No Lackin Challenge", which developed and is most prolific in the United States. The term "lackin" is a contraction of "lacking", which in the stunt references a "lack" of possession of a firearm and cautions viewers "not to lack" one. The stunt involves filming someone, or oneself, brandishing a firearm and pointing it at an unsuspecting victim with the goal of provoking them to withdraw a firearm, or otherwise be mocked for being found without, i.e., "lacking", the weapon. The game has been popularized as a way for people living in areas with high incidence of gun violence to test their readiness to protect themselves. However, the game is virtually unknown in Bangladesh and India, so the policy and its intended meaning of prohibiting violence and its advocacy may be lost in the Bengali translation. Moreover, the consequence of this lack of contextualization is that readers in Bengali may infer that they are not intended audience of the policy and that, for this reason, the policy does not apply to them.

The contextualization of policy stipulations regularly referenced resources and examples that are unfamiliar to most readers in Bangladesh and India. References to emergency phone numbers and organizations, e.g., the National Center for Missing and Exploited Children, refer to US-based organizations that serve only the American public. There are few, if any, references to Bangladeshi and Indian emergency numbers and allied civil society organizations. The consequence is that the policy provides few operational resources for platform users in Bangladesh and India.

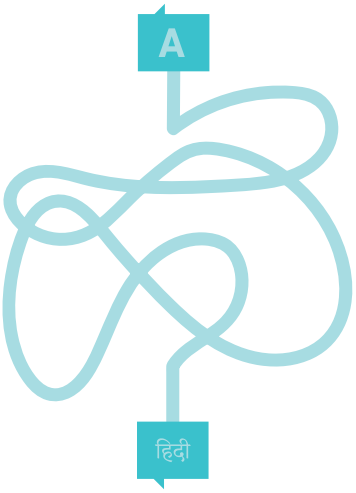
Table 9 Summary of YouTube Community Guidelines Review of Bengali Translation

Criteria	Finding	Example
Accuracy	Terms, ranging from technical to colloquial language, were regularly transliterated or translated ambiguously in Bengali.	Most of content appears to be in Bengali script, although terms may be transliterated, ranging from titles to terms like "হাইড্রক্সিক্লোরোকুইন" ("hā'iḍraksiklōrōku'ina") meaning "hydroxychloroquine".
Errors	There are grammatical and spelling errors of critical terms that need to be corrected for the policy to be correctly interpreted.	The term "suicide" is spelled incorrectly in the Bengali translation.
Clarity of meaning	The word for word translation from English to Bengali produced concatenations of each individually translated term that were technically accurate but semantically ambiguous or incoherent.	The translation of the phrase "what this policy means for you" is translated word for word as "এই নীতি আপনার জন্য মানে কি" ("Ē›i nīti āpanāra jan'ya mānē ki") when a better translation would adapt the language to say "এই নীতি আপনার জন্য কতটা অর্থবহ" ("ē›i nīti āpanāra jan'ya kataṭā arthabaha"), which translates more closely to "how meaningful this policy is to you."
Quality of expression	The inaccuracies in translation and the omissions of key terms in sentences limited the fluency and readability of the translation and signaled to translators that the texts were machine translated.	The quality of expression can read as unnatural and overly formal because the text includes words that few Bengali speakers use in everyday speech, and they reflect English expression in Bengali.
Consistency	Key terms from the English source texts were frequently omitted in Bengali translation.	Key terms like "report", "community", "creator" and "sign-in credentials", as well as numbers (1, 2, 3...) have not been translated from English and are simply omitted from the Bengali translation.
Diversity, equality, and inclusion	While the translation was based on Central Standard Bengali, the translation included terms specific to regions, dialects, and groups, which are not used by all or most Bengali speakers.	The translation of the term "official" as "আধিকারিক" ("Ādhikārika") is a term familiar to people in West Bengal, India, but not in Bangladesh. A more widely recognized term would be "দায়িত্বশীল কর্মকর্তা" ("Dāẏitbaśīla karmakartā"), translating roughly to "responsible officer".
User accessibility, register and tone	The translation adopted a formal register that requires expert reading proficiency in Bengali.	The vocabulary is often specialized or includes formal rather than colloquial language that is unfamiliar and not used by most speakers.
Contextualization	The policy explanations referred to examples most familiar to readers in Anglophone countries, especially Anglophone North America, and often included affiliate links in English.	The contextualization of "election misinformation" using examples of mis- and disinformation from Brazil and the United States requires an exceptional level of familiarity with global affairs and inference from readers when the content could be adapted to cite examples that are familiar to Bengali speakers in Bangladesh and India.

HINDI TRANSLATION REVIEW FINDINGS

Summary of Facebook Community Standards Review

Internews and Localization Lab's review of Facebook's Community Standards in Hindi showed that all 24 community standard policies were translated from English to Hindi and that the translated policies:

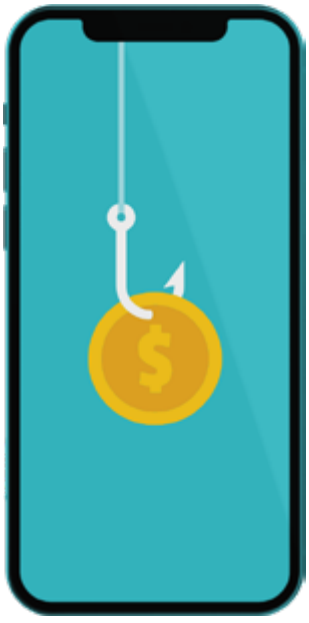


1. **Appeared to rely on machine translation to translate texts and produced texts with systematic inaccuracies in word choice, errors in grammar and spelling, and reliance on Hinglish, a macaronic hybrid use of English and Hindi;**

2. **Predominantly relied on verbatim, or word for word, translations from English that, while neutral in terms of dialect, requires a high level of technical knowledge to understand;**

3. **Consistently lacked socio-cultural contextualization to explain key policy concepts as they apply to users' experiences in India and South Asia more broadly, as social stratification and discrimination can look different in different cultural contexts.**

The review findings show that Facebook's Community Standards are of limited quality and usability to most Hindi speakers. This is due primarily to the inaccuracies in vocabulary and grammar and the reliance on English to translate, and frequently to transliterate, key terms into Hindi. These factors limited the quality of the translation because it created texts that had systematic errors, employed complex, difficult to follow sentence structure, and used terms that were highly technical and, at times, obsolete in Hindi. This impacted the usability of the text for all but some readers with an expert understanding of English and, specifically, information technology terms in English. The result is that the policies are inaccessible to most Hindi speakers and even less accessible to speakers for whom Hindi is a second or third language or to speakers who do not know technical terms in English. Such use of technical language creates a barrier for many end-users, many of whom are first-generation technology users and may not have knowledge of technical jargon in the policies. The use of machine translation contributed to a lack of coherence and clarity of meaning in the translations and, at times, the communication of ideas contrary to those in the source texts. Moreover, while the translations generally reflected the English language source texts, they often did not reflect spoken or written Hindi, the consequence of which were texts that read as obsolete, unnatural, and occasionally nonsensical in Hindi.



- For one example, terms like content, third party, scam, advance fees scam, or Ponzi or Pyramid scam have been transliterated into Hindi. This requires readers to have a high level of financial literacy in addition to proficiency in technical English. For many readers, concepts like "scams" or "Ponzi scheme" require explanation and contextualization in Hindi, and the omission of such explanations creates barriers to comprehension.

- For another example, the mistranslation of terms, e.g., conjunctions, can construe meanings contrary to what the original source text communicates. The policy prohibiting kidnapping, for instance, stated that content about kidnapping is prohibited if the content is not being used to advocate for a person's human rights or safe return home. This use of "if" is inaccurate, and a better translation would instead use the conjunction

"where" ("jahaan") in Hindi. While the conjunction "if" can often be used interchangeably with "where" in English, this is not the case in Hindi. Here, the use of "if" in the sentence about kidnapping content creates ambiguity by suggesting that human rights advocacy content may get moderated, while content promoting kidnapping may not.

These inaccuracies show that the text requires readers to have a strong understanding of both English and Hindi and to possess a high level of reading comprehension around technical and information technology terms. Even still, the at-times limited coherence requires readers to refer to the English original to fully grasp the policy.

Consistent with the other policy translations, the policies in Hindi lack sociocultural contextualization that can create ambiguity and clarity for readers in South Asia. For instance, the references to prohibited organizations, ideologies, and hate speech include "Nazism", "white supremacy", "white nationalism", and "white separatism", which refer to forms of hate, prejudice, and institutionalized inequalities that are most concentrated in North America and Europe. Rather, in the context of India, Hindu Rashtra, Hindutavawad, Saffronization, caste, and religion are concepts that are more reflective of social, economic, and political inequities and forms of hate that resonate with readers of Hindi. Including localized examples would better communicate the concept of the policy because it would provide concrete examples and explanations in a Hindi-language context for users in India and South Asia, primarily. This would leave less room for readers to make the oblique interpretation that derogatory and abusive speech most prevalent in India, caste-based anti-Dalit sentiment for example, are permitted on the platform.

Table 10 Summary of Facebook Community Standards Review of Hindi Translation

Criteria	Finding	Example
Accuracy	The text in Hindi frequently included terms that do not make sense or that did not deliver the meaning of the original policy to Hindi speakers.	The translation of terms like "law enforcement" as "कानून प्रवर्तन" ("kaanoon pravartan"), "disabled" as "अक्षम" ("aksham") and "public visibility" as "सार्वजनिक दृश्यता" ("saarvajanik drshyata") do not resemble the ways that Hindi speakers refer to these concepts.
Errors	The translated policies had numerous spelling and grammatical errors in Hindi.	The sentence "हम इस कंटेंट को ढँक देते हैं ताकि लोग चुन सकें कि उन्हें यह देखना है या नहीं" from the policy on violent and graphic content translates roughly to "We cover this content so people can choose whether or not to watch it". The translation took the literal meaning of "add a cover", i.e., a lid. This metaphoric use of "cover" requires explanation.
Clarity of meaning	The reliance on word for word translation decreased the semantic clarity of the policy's overall meaning.	The translated policy on incitement of violence states "जैसे आतंकवादियों को मार ही डालना चाहिए", which suggests that only direct violence is prohibited, rather than direct violence and incitement of violence. The boundaries between violence and incitement are blurred in translation.
Quality of expression	The translation reflected English syntax and, in particular, parenthetical phrases and dependent clauses rather than Hindi syntax.	The policy on bullying and harassment includes a list of prohibited actions and starts the list with an imperative verb "do not:" followed by infinitives that complete the verb phrase, e.g., "do not... post". This completion of the verb phrase is lost in the Hindi translation, where the infinitive verb is mistranslated as other parts of speech.
Consistency	The policy regularly translated the same English term differently in Hindi without an obvious semantic reason.	In some parts of the policy, the term "content" (i.e. user-generated matter that users share on the platform) has been translated as "सामग्री" ("saamagree"), which roughly translates to "materials", while in other instances, it has been transliterated as "कंटेंट", or "kantent"
Diversity, equality, and inclusion	The policy document included few cultural adaptations that would make the meaning of the policy resonate with readers in South Asia.	The examples of hate and political extremism cite examples from Western countries such as white supremacy, white nationalism, and white separatism instead of localized examples like Hindu Rashtra, Hindutavawad, Saffronisation, Caste, Religion etc.
User accessibility, register and tone	The translation heavily relied on English—Hinglish—words in most sentences to convey technical terms.	When one opens the links there are examples of how to report incidents and all of the options on the graphics are in English.
Contextualization	Most of the policies included explanation of the policies using references familiar to readers in Anglophone countries, as well as include affiliate links in English and referenced civil society and public organizations that serve the American public.	The examples of prohibited hate speech and incitement of violence make reference to hate groups like the KKK (Ku Klux Klan) and the American Nazi Party that do not operate in India.

Summary of YouTube Community Guidelines Review

Internews and Localization Lab's review of Facebook's Community Standards in Hindi showed that all 22 community standard policies were translated from English to Hindi and that the translated policies:

1. **Overall employed grammatical and correctly punctuated sentences in translated content that contributes to readers' ease of reading and high quality of expression;**

2. **Generally relied on Hinglish to communicate technical and key policy terms, which create barriers for readers unfamiliar with English-language technical terms to engage meaningfully with the policy content;**

3. **Inconsistently translated affiliate and additional content related to policies, which creates barriers for readers to engage with explanatory materials about the policy terms and conditions.**

The review findings indicate that YouTube's Community Guidelines, when translated, were generally readable and had high quality of expression from English to Hindi. These strengths were aided using short, direct sentences following a clear subject-verb-object structure that is more accessible to a wider range of readers.

While the translation's fluency in Hindi meant facilitates user engagement in Hindi, there are several areas of the translation that could be strengthened: these include the localization of technical terms into Hindi, rather than Hinglish, for readers in Hindi and, especially, for readers with varying levels of reading and informational literacy. For one, technical terms like "profile", "link", "site", and "website" have simply been transliterated into Hindi. While such terms may be used among some Hindi speakers, these types of terms may require additional explanation, e.g., through a glossary, for most readers. More fundamentally, this type of translation and use of technical jargon does not provide readers with culturally familiar terms, explanations, or definitions, and can consequently make meaningful engagement with and comprehension of the policies difficult.

> **What is Hinglish?**
> The term "Hinglish" is a portmanteau of "Hindi" and "English" and references to a fusional use of English with Hindi and other languages of the Indian subcontinent[24]. The hybrid involves code-switching and interpolating—also called "translanguaging"—Hindi or English words, phrases, and even whole sentences into on or the other language, as well as adapting and translating certain words for use in the other language.

Additionally, the community guidelines were inconsistently translated into Hindi and employed limited contextualization and localization for readers outside of Anglophone countries. For instance, in the policy on child safety, the page on "Determining if your content is "made for kids"' or how to "Age-restrict your video" remained in English, which means that users must use a separate resource like Google translate or rely on an English speaker to translate the document for them, both of which create a high barrier to engagement.

Similarly, there were also occasional, but not systematic, inconsistencies within translated texts, such as by leaving content in English. In the case of translating proper names like "YouTube" or "Google" or by including

affiliate links to English language content. For example, the link to the online internet safety course for youths called "Be Internet Awesome", it would be helpful to provide transliterations of names like Google and to provide localized translations of affiliate content. This would increase readers', particularly those with limited ability to read the Latin alphabet, ability to engage with the policy. For branded content like "Be Internet Awesome," which is outside the scope of the policy, it would be helpful to mark it as such and to provide a disclaimer that the content is not translated into Hindi if no localized translation exists. These types of cues will enable readers to better engage with the platform's content.

Table 11 Summary of YouTube Community Guidelines Review of Hindi Translation

Criteria	Finding	Example
Accuracy	The translation's reliance on transliteration to communicate key policy terms reduced the readability and the recognizability of the translation as "Hindi", as opposed to "Hinglish".	Terms terms like "प्रोफ़ाइल" ("profail"), "लिंक", ("link"), "साइट" ("sait"), and "वेबसाइटों" ("vebasaiton") have been transliterated into Devanagari script without conceptualization, definition or usage, at times, of Hindi terms that people use to denote these concepts.
Errors	Most content was written in grammatically correct, correctly punctuated, and correctly spelled Hindi, although there were occasional copy errors.	Occasionally words and phrases, such as the phrase in Hindi "domestic", written as "घरेलु" ("gharelu") instead of "घरेलू" ("ghareloo"), were mispelled.
Clarity of meaning	The correct use of grammar, punctuation and spelling significantly increased the readability of the text even when terms or English acronyms were ambiguous.	English abbreviations have been used frequently in the text, such as "Facebook", "SSN", "PPS", "ITIN", "GPS", "Google Maps", etc.
Quality of expression	The translation was generally easy to follow because the translation had few errors in grammar, punctuation, and spelling, although some translations were overly literal, which creates semantic ambiguity.	The translation of putting a cover, or a screen, over a distressing image was translated literally as adding a cover (i.e., a lid), which for Hindi speakers is ambiguous and needs further contextualization.
Consistency	The translation mostly includes correct Hindi grammar, with barriers to comprehension primarily stemming from translation and transliteration.	In some places the text translates the term "content" as "सामग्री" ("saamagree"), roughly translating to "material", and in other places the text uses a transliteration as "कंटेंट" ("kantent").
Diversity, equality, and inclusion	The use of Hinglish and transliterated English in Devanagari script decreased the readability and required readers to have a high reading level in both Hindi and English.	People who are new to such platforms, people who have limited literacy in Hindi, and who are not familiar with English language terminology, as well as people who speak Hindi as a second or third language may find the texts very difficult to read and comprehend.
User accessibility, register and tone	While the use of Hinglish could create ambiguity in meaning, its main benefit is that it contributed to a casual, colloquial tone that was more reflected of natural language.	Many end-users of the policy are new to using the Internet and need policies to be written in ways that gives them confidence that they can read the document without difficulty.
Context	Nearly all the affiliate content was included in Hindi, with only a few external links to videos in English.	All the external links have content available in Hindi. An affiliate video in the policy on spam, deceptive practices, and scams policies was included in English but has Hindi subtitles with frequent use of Hinglish.

RECOMMENDATIONS

Internews and Localization Lab's recommendations for improving the quality and usability of the content moderation policy translations are informed both by the findings of this review and by best practices in translation. Improving content moderation globally starts with making content moderation policies intelligible and relevant so that end-users can identify with the policies and relate to them. It is thus critical to adjust the policies to local contexts, such as by explaining prohibited content like hate speech in ways that are relevant to end-users in each language. Without cultural adaptation, the policies lose social and psychological relevance to end-users, which inhibits the ability of social media corporations to build and moderate spaces that protect freedom of expression, while simultaneously promoting safety and respect for all communities online.

Recommendations for Localization Policy

A Companies should publicly commit to a review of existing policy translations and processes.

To address the serious flaws with current policy translations Meta and Google should commit to a full review of existing translations, as well as their wider translation and localization processes. Based on the reviews of the languages covered by this report it is likely that other languages suffer from the same systemic issues, and a complete review is required. This likely applies to other platforms and companies not covered by this report, and a framework for best practice translation would benefit the sector.

B Companies should commit to translating policies into all languages in which their products are available.

While errors in policy translations can lead to harmful confusion and deny users agency, there are many languages for which both Facebook and YouTube have localized their products without providing any policy translations at all. It is estimated that this impacts hundreds of millions of platform end-users, including in contexts with particularly high risks of harm. Translation of public facing policies should be an automatic step whenever any product is localized in a new language.

Recommendations for Localization Processes

C Define the goal of the policy in translation.

Articulating the goal of the policy in translation will help to provide objectives to reach in the completed translation and serve as a basis for setting indicators and outcomes for translation monitoring and evaluation.

D Select and follow a method of translation best practice.

Choosing a method of translation is an important step in the translation process and helps to increase quality and replicability of similar quality translations going forward. Common best practices include:

- Creating the policy separately in each language
- Conducting a one-way translation from the source into the target language

- Conducting a two-way or "back" translation from the source into the target language and then back into the source language

While any of these methods can facilitate quality translations, Localization Lab suggests that Facebook and YouTube use two-way translation because it most efficiently enables multilingual teams to conduct quality reviews and to ensure that the translation accurately communicates the source language of the policy.

E Identify the target audience of the policy in translation.

Ascertain what are the socio-demographic characteristics of the average user in each language cohort and define the target audience as widely and inclusively as possible.

F Determine the language standard or dialect in which to translate the policy.

For languages with regional variation, identify a standard language in which to translate the text or consider translating the document into multiple dialects. Likewise, it is important to note that certain language choices may also be political choices and to demonstrate the organization's tangible commitment to inclusion of marginalized communities.

G Establish the target level of readability and translate to that level of readability.

Creating internationalized targets for readability will increase users' ability to access, comprehend, and engage fully with the translated policies. Localization Lab recommends that the policies are written to match as best as possible the reading level of the average adult in the target audience. Increasing readability involves several steps, which include:

a Understanding the reading level of the average adult in the target language.
b Breaking up long sentences in favor of short sentences.
c Using straightforward, simple syntax and verbs.
d Reducing jargon, e.g., technical vocabulary and creating glossaries with definitions.
e Checking spelling and grammar.
f Writing in a conversational tone.

> ! Including glossaries of key technical terms in each language will increase the usability of the policy with readers of varying levels of reading literacy and technical understanding.

H Adjust policies to context and incorporate relevant localized examples to explain key concepts and terms.

Adapting the examples in the policies helps readers to see themselves in the policies and to relate to them, which ultimately enable the policies to resonate with the target audience.

For example, when contextualizing how, why, what discriminatory language is prohibited, in a Hindi language context it may make sense to talk about caste and Saffronization rather than racism and white nationalism. This type of reframing enables readers to quickly grasp what the policy means.

I Translate the texts for meaning.

Good translation requires linguistic and cultural adaptation in the target language so that the documents are relevant to the target audience. Translating phrases rather than translating word for word and adapting and defining examples of prohibited behaviors to examples relevant to users' local contexts will improve the quality of the translation.

J Review the translated policy for quality and usability and gain feedback from representative samples of end-users in each language community.

Proofread and copy edit the translation for errors and omissions for quality control. Obtain end-user feedback on the translation's quality, usability, and bias from a representative sample of end-users in each language community. Moreover, it is important to involve different reviewers than the translators, as this improves and ensures a higher quality translation.

K Establish translation quality control and quality assurance mechanisms and procedures that prioritize quality and usabilWity of translation and that translate and localize texts in partnership with end-user communities.

Producing high quality translations of content moderation policies is the first step in building social media ecosystems that respect all users. This includes developing translation guidelines and translation review checklists for each translation. Such reviews need to cover aspects that include, but are not limited to accuracy, grammar, style, spelling errors, linguistic standardization (i.e., omitting use of idioms and colloquialisms, such as "that's a piece of cake"), social bias (i.e. identifying and omitting biased, discriminatory, and offensive language), and readability. It is also important to establish review mechanisms, such as internal reviews, user-feedback on beta versions, manager sign off, frequent independent audits by local experts, and ongoing engagement and feedback from end-user communities to ensure that platforms publish high quality translations that clearly and appropriately communicate the policy terms to users in each language of the policy.

CONCLUSION

The findings of this review show that there are numerous translation barriers that inhibit end-users from meaningfully engaging with Facebook's Community Standards and YouTube's Community Guidelines. These barriers range from issues of accuracy to cultural relevance to readability. Key terms were regularly mistranslated, resulting in ambiguous, illogical, and incoherent statements in the target languages. These errors mean that the translated policies regularly communicate ideas contrary to what the policies state in English. Such errors can create policy loopholes that do not exist in the English original, such as the Arabic translation of the statement that Facebook may censure calls to violence as Facebook censuring telephone calls to violence. The result is that readers may think that only telephone calls are prohibited due to the serious inaccuracies in the translated policy or even that Facebook audio calls are surveilled.

While this study was limited in scope to four languages and two reviewers per language, the findings suggest that many of the translation issues, in particular mistranslations and readability for the platforms' diverse audiences, are systematic and recurrent within a translated set of policies in one language and across languages. The findings also show that while the translations on Facebook and YouTube both share significant issues of contextualization and logical coherence, there are distinctions between the types of translation issues they present: namely, the issues on Facebook are frequently ones of cultural sensitivity, while the issues on YouTube are often ones of missing text, numbers, and sentences, which results in an incomplete translation.

These findings point to the relevance of working directly with communities to improve both the accuracy and the relevance of the translation: to understand not only how best to translate a term and communicate key concepts, but also to work with communities to create new terms to communicate concepts from English that may have no equivalent term yet in the target language. This type of collaboration with end-users is a critical step in advancing a baseline understanding about expectations of engagement.

While high quality translations are a prerequisite to improving content moderation, they also open questions about the target audience of the policies and how platforms will moderate content. Translating the content so that end-users in each language community can meaningfully engage with the policies will foster the reasonable expectation that prohibited content will be moderated and removed, when historically this has not always been.

Addressing such issues is urgent because they bear upon the lives of billions of people globally and the risks of harm can be high for the safety, security, and protection of end-users. Consequently, these questions urge social media platforms to shift their focus in content moderation toward advancing common standards of translation; common standards of understanding; and common standards of moderation.

ENDNOTES

1 Wikipedia, last edited Sept. 20, 2022. List of social media platforms with at least 100 million active monthly users. Accessed October 17, 2022 https://en.wikipedia.org/wiki/List_of_social_platforms_with_at_least_100_million_active_users

2 Langvardt, Kyle, 2018. Regulating Online Content Moderation. The Georgetown Law Journal. 106(5). Accessed October 17, 2022. https://www.law.georgetown.edu/georgetown-law-journal/in-print/volume-106/volume-106-issue-5-june-2018/regulating-online-content-moderation/

3 Gillespie, Tarleton. August 20, 2021. Content moderation, AI, and the question of scale. Big Data & Society. 7(2). Accessed October 17, 2022. https://journals.sagepub.com/doi/full/10.1177/2053951720943234

4 Fick, Maggie and Paresh, Dave. April 23, 2019. Facebook's Flood of Languages Leave It Struggling to Monitor Content. Reuters. Accessed October 17, 2022. https://www.reuters.com/article/us-facebook-languages-insight/facebooks-flood-of-languages-leaves-it-struggling-to-monitor-content-idUSKCN1RZ0DW

5 Ibid.

6 Meta Investor Relations, October 26, 2022. "Meta Reports Third Quarter 2022 Results", Press Release. https://investor.fb.com/investor-news/press-release-details/2022/Meta-Reports-Third-Quarter-2022-Results/default.aspx

7 Reportal, August 15, 2022. "YouTube Statistics and Trends" https://datareportal.com/essential-youtube-stats

8 Statista, October 2022. "Distribution of Facebook Users in the United Kingdom as of September 2022, by agegroup". Accessed November 9, 2022. https://www.statista.com/statistics/1030055/facebook-users-united-kingdom/

Datareportal, August 15, 2022. "Facebook Statistics and Trends". Accessed November 9, 2022. https://datareportal.com/essential-facebook-stats?rq=Facebook

Statista, July 13, 2022. "Share of Facebook users in Canada as of June 2022, by age group". Accessed November 9, 2022. https://www.statista.com/statistics/863754/facebook-user-share-in-canada-by-age/

NapoleonCat. n.d., "Facebook users in Ireland." Accessed November 9, 2022. https://napoleoncat.com/stats/facebook-users-in-ireland/2021/01/

Statista, July 27, 2022, "Number of Facebook users in Australia from 2015 to 2022". Accessed November 9, 2022. https://www.statista.com/statistics/304862/number-of-facebook-users-in-australia/

9 Statista, July 20, 2021. "Forecase in the Number of YouTube users in the United Kingdom from 2017-2025". Accessed November 9, 2022. https://www.statista.com/forecasts/1145489/youtube-users-in-the-united-kingdom

Algonquin College, n.d., "YouTube Stats". Accessed November 9, 2022. https://www.algonquincollege.com/ac-social-media/youtube-stats/

Correll, D. April 1, 2022. "Social Media Statistics Australia—March 2022". Accessed November 9, 2022. https://www.socialmedianews.com.au/social-media-statistics-australia-march-2022/

Datareportal, February 15, 2022. "Digital 2022: New Zealand", Accessed November 9, 2022. https://datareportal.com/reports/digital-2022-new-zealand

Datareportal, August 15, 2022. "YouTube Statistics and Trends". Accessed November 9, 2022. https://datareportal.com/essential-youtube-stats

Datareportal, February 15, 2022. "Digital 2022: Ireland". Accessed November 9, 2022. https://datareportal.com/reports/digital-2022-ireland

10 Meta, n.d., Facebook Community Standards. Facebook Transparency Center. https://transparency.fb.com/policies/community-standards/;

Facebook., Homepage, Accessed October 17, 2022. https://www.facebook.com/ ;

YouTube, Homepage, Accessed October 17, 2022. https://www.youtube.com/;

YouTube, n.d., YouTube Community Guidelines. Accessed October 17, 2022. https://www.youtube.com/howyoutubeworks/policies/community-guidelines/?gclid=Cj0KCQjw166aBhDEARIsAMEyZh5HvKp8NuBlhZB3QPstyuNOxlzY0uwnl-m3zziWdNYhAXjrBlAq6rkaAkUUEALw_wcB

Twitter, Homepage, Accessed October 17, 2022. https://twitter.com/;

Twitter, Rules and Policies, Accessed October 17, 2022. https://help.twitter.com/en/rules-and-policies/twitter-rules

11 Culliford, E., and Heath, B., October 26, 2021. Language Gaps in Facebook's Content Moderation System Allowed

Abusive Posts on Platform: Report. The Wire, India. Accessed November 9, 2022. https://thewire.in/tech/facebook-content-moderation-language-gap-abusive-posts

12 The Washington Post, September 2021. "The Facebook Files". Accessed November 9, 2022. https://www.wsj.com/articles/the-facebook-files-11631713039?mod=svg-breadcrumb

13 Global Freedom of Expression, Columbia University, "The Gambia v. Facebook". Accessed November 9, 2022. https://globalfreedomofexpression.columbia.edu/cases/gambia-v-facebook/

14 Oremus, W. August 25, 2021, "'YouTube magic dust': How America's second-largest social platform ducks controversies". The Washington Post.

15 Simonite, T. October 25, 2022. "Facebook is Everywhere; It's Moderation is Nowhere Close", Wired. Accessed November 9, 2022. https://www.wired.com/story/facebooks-global-reach-exceeds-linguistic-grasp/

16 Facebook Help Center, "How to Report Things". Accessed November 9, 2022 https://www.facebook.com/help/1380418588640631?helpref=hc_fnav

17 YouTube Help". Accessed November 9, 2022. https://support.google.com/youtube/answer/2802027?hl=en&co=GENIE.Platform%3DDesktop

18 Datareportal, August 15, 2022, "Facebook Statistics and Trends". Accessed November 9, 2022: https://datareportal.com/essential-facebook-stats?rq=Facebook

19 Datareportal, August 15, 2022, "YouTube Statistics and Trends". Accessed November 9, 2022. https://datareportal.com/essential-youtube-stats

20 Levine, M. et al. ABC News, "Nation's deadliest domestic terrorist inspiring new generations of hate-filled 'monsters', FBI records show". Accessed November 9, 2022. https://abcnews.go.com/amp/US/nations-deadliest-domestic-terrorist-inspiring-generation-hate-filled/story?id=73431262

21 Fatafta, M. November 18, 2021. "Facebook is bad at moderating in English. In Arabic, it's a disaster", Rest of World.
Accessed November 9, 2022. https://restofworld.org/2021/facebook-is-bad-at-moderating-in-english-in-arabic-its-a-disaster/

22 Abid, Abubakar, Farooqi, Maheen, Zou, James. July 2021. Persistent Anti-Muslim Bias in Large Language Models. AAAI/ACM Conference on AI, Ethics, and Society. Accessed November 9, 2022: https://arxiv.org/abs/2101.05783

23 Abid, Abubakar, Farooqi, Maheen, Zou, James. July 2021. Persistent Anti-Muslim Bi-as in Large Language Models. AAAI/ACM Conference on AI, Ethics, and Society. Accessed November 9, 2022: https://arxiv.org/abs/2101.05783

24 Kothari, R., and Snell, R. 2012, Chutneyfing English: The Phenomenon of Hinglish. Penguin Global.

ACKNOWLEDGMENT

This report is the result of a collaboration between Internews and Localization Lab that began in early 2022 on evaluating the quality and usability of social media content moderation policies in translation. The tandem goals of the project include advancing the technology industry toward commonly accepted minimum standards of translation and creating standards that improve the quality and usability of translated social media policies.

The team would like to express their gratitude to all the people who made this project possible. We are grateful to the translators who generously shared their perspectives and insights. Without their input, this review would not have been possible. We are likewise grateful to Mustapha Al-Abdali for the graphic design of this report.

The team also extends its appreciation for the dedication of the Internews and Localization Lab teams. The vision and financial support of Rafiq Copeland and Zoey Tung Bartholomey of Internews made this project possible. At Localization Lab, our appreciation goes to Muna Hemoudi for led the research design and data collection, Orla O'Sullivan led data analysis and report writing, Dragana Kaurin and Giulia Balestra for envisioning and advancing the project's direction and future.

The team hopes that the findings of this study will be used by content moderation experts, translation experts, community members and all parties interested in advancing equity, safety, and trust on social media.

Contact

Orla O'Sullivan
Research Associate
Localization Lab
e: orla@localizationlab.org

Dragana Kaurin
Executive Director
Localization Lab
e: dkaurin@localizationlab.org

Rafiq Copeland
Senior Advisor
Internews
e: rcopeland@internews.org

LOCALIZATION LAB

Internews
Local voices. Global change.