A report from

cdt | Research

Moderating Maghrebi Arabic Content on Social Media

Mona Elswah

September 2024

CENTER FOR
DEMOCRACY
& TECHNOLOGY

The Center for Democracy & Technology (CDT) is the leading nonpartisan, nonprofit organization fighting to advance civil rights and civil liberties in the digital age. We shape technology policy, governance, and design with a focus on equity and democratic values. Established in 1994, CDT has been a trusted advocate for digital rights since the earliest days of the internet. The organization is headquartered in Washington, D.C., and has a Europe Office in Brussels, Belgium.

MONA ELSWAH

Author

A report from

cdt | Research

Moderating Maghrebi Arabic Content on Social Media

CDT Research Report

Mona Elswah

WITH CONTRIBUTIONS BY

Samir Jain, Dhanaraj Thakur, Aliya Bhatia, and DeVan L. Hankerson.
Illustrations and additional design by Osheen Siva.

ACKNOWLEDGMENTS

We would like to thank the team of Digital Citizenship, including Emna Mizouni, Wiem Ouardani, and Aly Mhenni, for their invaluable support. We thank the members of our Advisory Committee; Afef Abrougui, Juan Carlos Lara, Nanjira Sambuli and Jillian York. We also thank the study participants who generously shared their time and insights with us. We thank Nathalie Maréchal for graciously translating the survey instrument. We also gratefully acknowledge the feedback and suggestions of Kate Ruane, Michal Luria, Gabriel Nicholas, and Farhana Shahid. Lastly, this work could not have been completed without the support of Timothy Hoagland, George Slover and Sarah Zolad. All views in this report are those of CDT.

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

Suggested Citation: Elswah, M. (2024). Moderating Maghrebi Arabic Content on Social Media. Center for Democracy & Technology. https://cdt.org/insights/moderating-maghrebi-arabic-content-on-social-media/

References in this report include original links as well as links archived and shortened by the Perma.cc service. The Perma.cc links also contain information on the date of retrieval and archive.

Contents

Introduction

People express their deepest desires, emotions, and imaginative ideas through language. However, Global North languages, particularly English, dominate in knowledge-sharing and technology. This has led to the marginalization of, and inadequate support for, Global South languages in digital spaces, particularly in the realm of content moderation. Content moderation involves governance mechanisms that determine participation in a certain community and control what is seen and what is valued online (Grimmelmann, 2015). By removing or reducing harmful content, content moderation helps ensure positive user experiences, safeguards social media platforms' brand reputation, enables compliance with legal requirements, and increases advertising revenues (Roberts, 2018). Initially reliant on human moderators, the vast scale of online content posted online necessitates integrating automated moderation systems. These systems, however, often struggle with Global South languages due to inadequate and unrepresentative training data in addition to a lack of understanding of the cultural nuances that inform the meaning of language.

In this report, part of a CDT series investigating content moderation biases and disparities in the Global South (Elswah, 2024), we specifically examine the challenges and implications for moderation of content in the Maghrebi Arabic dialects in North Africa. Additionally, the Maghrebi Arabic dialects are under-examined in the literature of content moderation. Using a mixed-method approach combining interviews, focus groups, and online surveys, we found that:

Most US-based social media companies utilize a global content moderation strategy that applies the same policies worldwide, whereas TikTok adopts a localized approach with many of their policies tailored to each region, particularly regarding cultural matters.

Maghrebi Arabic users employ tactics such as "algospeak" to evade moderation algorithms, because they believe they are being censored for political reasons. Many of them compensate for ineffective reporting mechanisms by using mass reporting to remove harassing content.

The lack of diversity in natural language processing teams that develop automated content moderation systems at social media companies, combined with insufficient training datasets for Maghrebi Arabic dialects and the recruitment of non-native annotators, negatively impacts the accuracy of the automated content moderation process.

Content moderators, who work under harsh conditions, are assigned content from any country in the Arab world and are expected to make decisions despite the challenging cultural and linguistic nuances that vary across the region. This can lead to errors in content moderation in some instances.

The Maghrebi Arabic Dialects: The Post-Colonial Legacies

Arabic is one of the most widely spoken languages in the world. It is also a very expressive language, with many dialects and variations in which words can carry different meanings and emotions. Across the Arab world, more than 400 million people in 22 countries speak this language in various dialects (Guellil et al., 2021). Arabic has evolved over centuries through interactions with Ottoman, British, French, and Italian rulers and various other cultures and civilizations. Geographically, dialectical Arabic is divided into two main categories: Middle Eastern (Mashriq) and North African (Maghreb) dialects[1] (Harrat et al., 2018). This report focuses on the Maghrebi Arabic dialects (or Maghrebi Darija, دارجة) in Morocco, Tunisia, and Algeria, which share numerous socio-historical and cultural similarities and have common linguistic specificities (Harrat et al., 2018).

The Maghrebi Arabic dialects are colloquially spoken Arabic, with each dialect having a unique set of vocabulary and nuances. Some words and phrases may hold different meanings or connotations across the Maghreb region. For example, the word قرعة (Qar'a) means a bottle of water in Algeria and a pumpkin in Tunisia. The Maghrebi Arabic dialects are influenced by many other languages, including Tamazight and other Amazigh languages, French, Spanish, and English.

In addition to its impact on the language itself, colonization, in combination with the shortcomings of available hardware and software, have influenced the way people write Arabic, leading to the emergence of a new form of writing in Arabic known as "Arabizi" (traditionally written as 3rabizi). Arabizi involves writing Arabic words using numbers and Latin characters in transliterated form (Darwish, 2013; Duwairi et al., 2016). It can be used to represent any Arabic dialect and is primarily used on social media platforms and in informal texting (Darwish, 2014; Hajbi et al., 2022). Arabizi has emerged as a response to the necessity of writing Arabic on platforms that lack native support for Arabic script, as well as the fact that many Arabic speakers are bilingual (Darwish, 2014). Another post-colonial effect on the Maghrebi Arabic dialects is what is known as “code-switching,” the use of two or more languages in the same sentence (Hajbi et al., 2022). This type of code-switching introduces an additional challenge, as it requires reading from left to right for the part written in Latin letters and in the opposite direction for the Arabic part (Younes et al., 2020).

Despite being spoken by over 100 million people, Maghrebi Arabic dialects are considered "low-resource languages." (Younes et al., 2020). This term refers to languages that have limited high-quality text data available, complicating the training of automated content moderation systems (Nicholas & Bhatia, 2023). The prevalence of Arabizi and code-switching further exacerbate these challenges, hindering effective automated moderation and the development of robust AI models. These challenges and more are further discussed in this report.

[1] There are other ways to categorize dialectical Arabic. For example, some divide it into 5 categories: Arabian Peninsula dialects, Egyptian, Levantine, Mesopotamian dialects, and Maghrebi Arabic dialects (Younes et al., 2020).

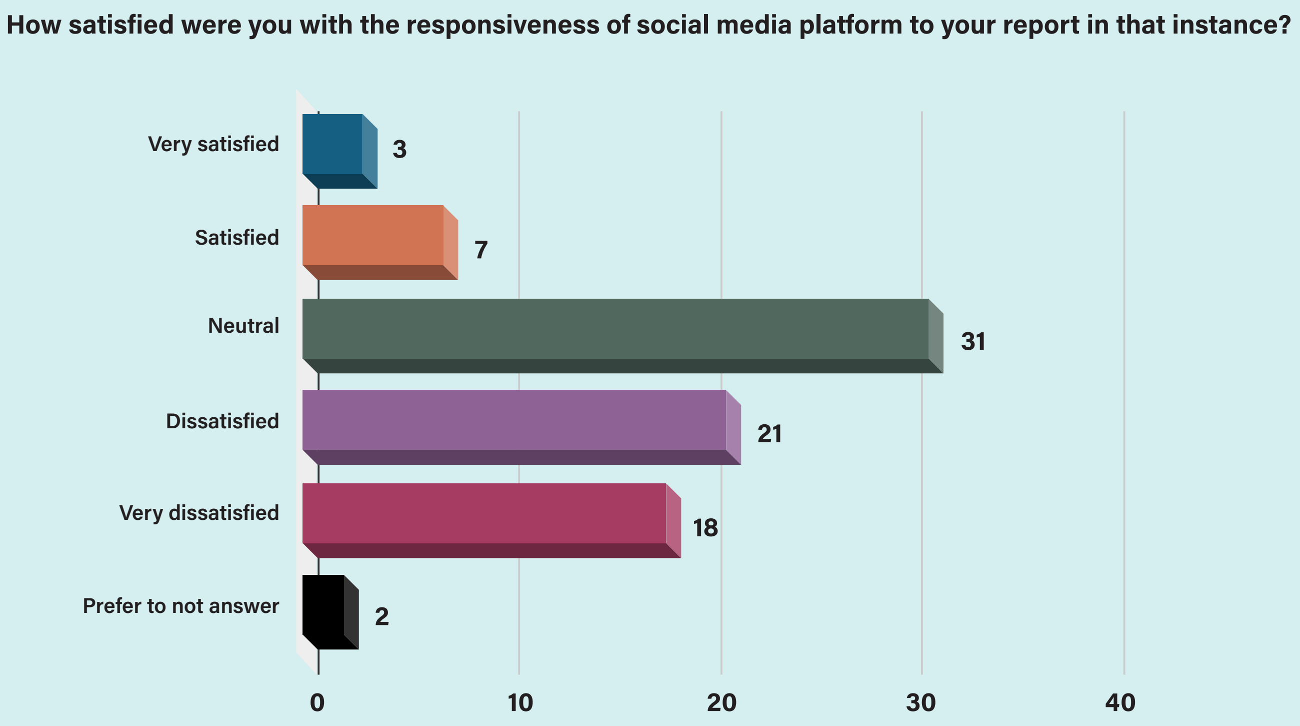
Main Findings

In this report, we examine how online services moderate Maghrebi Arabic dialects. We conducted 14 online interviews with content moderators and current and former employees from the Trust and Safety (T&S) and Policy departments at Meta, TikTok, YouTube, and X. We also interviewed digital rights advocates and conducted an online survey with 111 users of online services in Tunisia, Algeria, and Morocco. Additionally, we conducted five online focus group sessions with frequent Internet users and content creators in the Maghreb region (*see the Appendix for more details on the research methods used for this report*). Additionally, we organized a roundtable with Natural Language Processing (NLP) researchers from the Maghreb region to discuss the challenges and opportunities for developing reliable and fair AI models for Maghrebi Arabic dialects. **Based on this research, we made the following findings:**

1. Distrust of social media platforms in the Maghreb Region

The online public sphere in the Maghreb region is vibrant. Users rely on social media platforms to communicate, work, and engage in activism and political discussions. From our online survey, Facebook, Instagram, YouTube, and TikTok are the four most popular online services among Maghrebi Arabic-speaking users. **Despite this engagement, frequent social media users in the Maghreb region exhibit a level of distrust towards social media platforms.** About 57% of users who participated in our survey reported that they do not believe social media companies are doing their best to moderate content from the Maghreb region. About 58% expressed that they do not trust social media companies with their content. In addition, more than 62% of survey participants were concerned about their posts being removed. During the focus group sessions, participants repeatedly stated that they “think twice before creating any content" to avoid any takedown or restrictions.

This distrust may stem from the deeply rooted post-colonial perspective of people in the Maghreb region, reflecting historical sensitivities and power dynamics. The focus group participants expressed that the US-based origins of many social media companies contribute to their perception as distant entities detached from the local realities of users in the Maghreb region. This might also explain the growing popularity of



▲ **Figure 1.** Users' satisfaction with the responsiveness of social media platforms to the reporting process (% of those who said they reported content to a social media platform, n=82). Source: CDT's online survey (April - May2024).

TikTok in the region, as participants in the focus groups and interviews expressed a higher level of trust toward TikTok, owned by the Chinese company ByteDance, compared to American-owned Meta when it comes to expressing their views freely. A content creator from Morocco said:

> *"I think that anything you find on TikTok, you can also find on Instagram. The attraction of TikTok lies in its freedom. We notice a lot of shadowbanning on Instagram and Facebook, but TikTok is more free, which attracts many people who can consume content avidly for hours." (Content creator, Morocco, April 2024)*

This distrust is not just related to keeping content up, but also related to how companies respond to reports of content that users perceive as inappropriate. More than 74% of the survey participants said that they had previously reported content that violated the policies of some of the platforms they used. However, about 50% were dissatisfied with the responsiveness of the social media platform to their reporting, mentioning that the violating content remained on the platform (see Figure 1). A Tunisian content creator stated: "There is no transparent technology of reporting, nor an understanding of Tunisian, Arab, and local contexts."

This distrust, coupled with the absence of effective communication channels, has prompted some participants to take matters into their own hands, particularly when it comes to reporting content. Participants mentioned that they resorted to "mass reporting campaigns" to take down content or accounts that the social media companies have failed to remove using standard reporting channels. Mass reporting constitutes “abusing the reporting feature of social media platforms to delete accounts and pages” (Elswah, 2023). This tactic has been employed by some users in the Maghreb region who perceived social media companies as unfairly under-moderating content they found harassing.

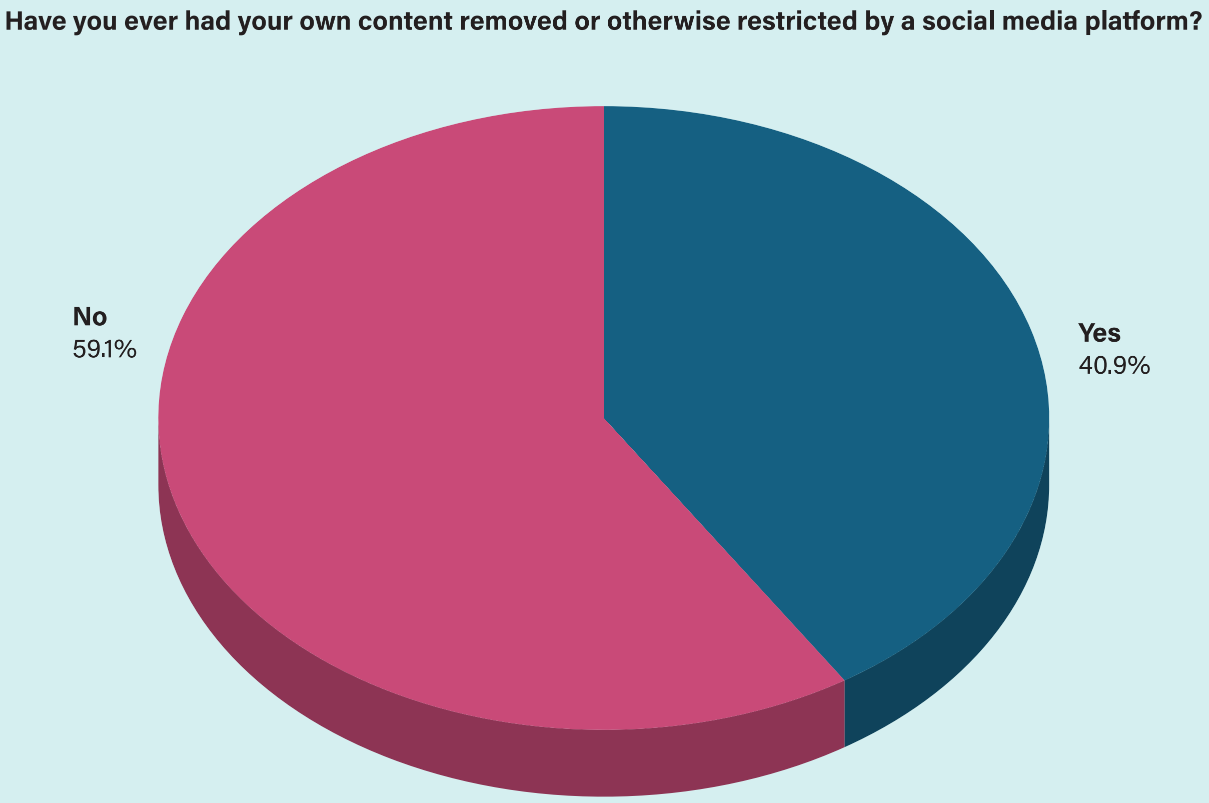
Consequently, civil society organizations in the region have taken on the responsibility of acting as the “middle man,” connecting concerned users with tech companies, especially in relation to takedowns and restriction of harmful content. This practice is now standardized, and responding to these reports is institutionalized within tech companies. However, digital rights advocates indicated in our interviews that tech companies may take months to respond even to this civil society intervention. A policy team member at a US-based social media company talked about this third-party escalation, stating:

> *“There are cases where things come to me through civil society organizations, for example, or through academics. I will escalate it internally for review. So, it is sometimes part of the process, too. It is like a kind of third-party escalation.” (Policy team employee, April 2024).*

This civil society escalation due to inadequate responses to reports from users underscores the need for more transparent and accessible communication channels for reporting and appeals.

2. Unexplained Content Removals and Shadowbanning

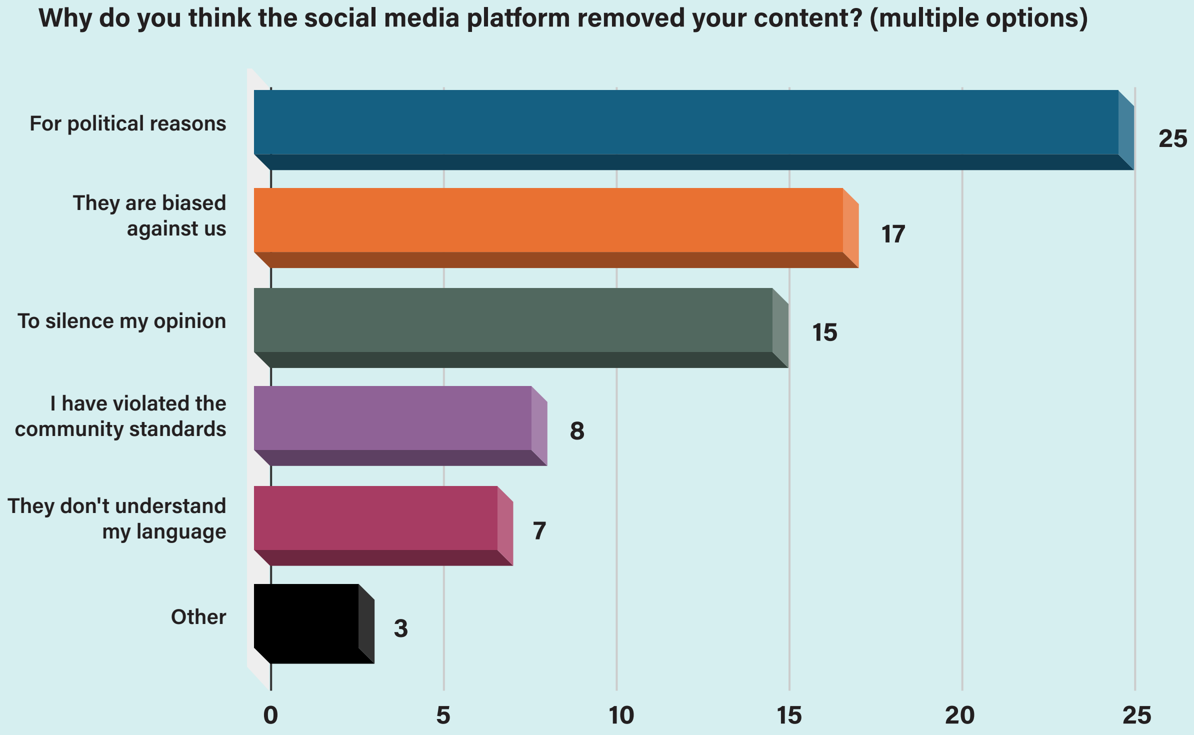
Participants in this study have experienced incidents of content removals or shadowbanning while posting content on various online platforms. Shadowbanning is a form of undisclosed content moderation process where a user's content is hidden or its visibility is significantly reduced without informing the user (Nicholas, 2022). Shadowbanning can make a user's comments, handle, or content appear normal to them but be invisible or difficult for others to find (Nicholas, 2022).



▲ **Figure 2.** Frequency of content removals or restrictions experienced by participants on different social media platforms (% of participants in the survey, n=110) Source: CDT's online survey (April-May 2024).

In the survey, about 40% of participants indicated that they have experienced content removal at some point, with the majority believing it was for political reasons (see Figures 2 & 3). Additionally, more than 62% of Maghrebi Arabic speakers who participated in the survey noted that they were concerned about being somehow silenced by social media companies.

The process of content removal is often confusing and unpredictable. Participants reported uncertainty regarding the timing and criteria for content removal, complicating their ability to anticipate such actions. During focus group sessions, some participants observed that seemingly harmless content is sometimes removed, while problematic content may remain unaddressed. Participants found shadowbanning—particularly of Palestine-related content—to be even more confusing.



▲ **Figure 3.** Users-perceived reasons for content removal. (Number of those who said that they have had their content removed or restricted by a social media platform, n= 82. Participants were allowed to select multiple options.) Source: CDT's online survey (April - May 2024)

Most participants cited the war in Gaza as precipitating the heaviest censorship they have encountered on social media platforms, particularly on Facebook and Instagram, both owned by Meta. According to participants, this feeling of being shadowbanned is prevalent with content related to Palestine in general. Content creators noted that their content about Palestine did not reach the same level of viewership as other content. An Instagram influencer noted:

> *"Whenever I mention Palestine in my stories, my stories do not get viewed. I usually get 20,000 people watching my stories, but when I talk about Palestine, I barely reach 1,000 views." (Content creator, Morocco, March 2024).*

Shadowbanning is a broad concept; all platforms design their algorithms to prioritize some content and deprioritize other content. In some cases, content moderators can choose to selectively demote certain content instead of removing it entirely. A TikTok moderator we interviewed revealed that one tool they have at their disposal is to classify a video as "hard to find." This manual tag would cause the algorithm to downgrade content, making it impossible for users to find the video in their searches.

In our interviews, a former Meta policy employee explained that this phenomenon may manifest through two primary mechanisms. First, shadowbanning can target certain "accounts" that frequently violate platform policies. Secondly, shadowbanning can target specific "topics or content areas." This is known as algorithmic deprioritization, especially in search or recommender systems (Nicholas, 2022). For example, Meta recently announced that they would stop proactively recommending political content on Instagram and Threads (Lorenz & Nix, 2024).

To avoid shadowbanning and content removals, some users and content creators in the Maghreb region have adopted what is known as "algospeak" tactics. Algospeak is the adoption of creative code words or phrases to avoid content removal or down-ranking by moderation systems (Lorenz, 2022). Participants in the focus group sessions and interviews highlighted four tactics of algospeak. Firstly, users strategically upload random content alongside material they believe will be restricted. Secondly, users manipulate their Arabic language by substituting letters with numbers or employing other unconventional spelling techniques like dotless Arabic. For example, instead of writing جمعة مباركة (translated as "blessed Friday"), a user would write it as حـــمـــعـه مـــٮـــاركــه, removing the dots and inserting additional spaces without changing the meaning. Thirdly, the use of symbols or emojis as substitutions for other concepts is also a common practice; one key example is the use of the watermelon emoji as a reference to the Palestinian flag. Lastly, some individuals opt to avoid certain words altogether.

Insights from interviews with NLP researchers and tech company employees revealed that these algospeak tactics may be successful in circumventing social media detection. However, platforms continuously update their algorithms through recurring training. Thus, one of the NLP researchers who worked at a tech company emphasized that "users need to continue being creative in their algospeak so that algospeak does not become something that the models can get used to and then detect."

3. The State Of Maghrebi Arabic Content Moderation

A. THE TRAUMA OF CONTENT MODERATION

As with other languages, content moderation for Maghrebi Arabic dialects is conducted using a combination of human and automated content moderation. Human (manual) content moderation was the first type of moderation used to manage content before the integration of automated tools and machine-learning models. A former T&S employee described the first years of content moderation at Meta:

> *"The plan had been basically to hire college students to moderate the other college students...But it [Facebook] kept growing, and it expanded first outside of colleges and then very rapidly outside of the United States. ... that meant we had to add more and more people to the moderation workforce." (Former Employee, Meta, April 2024).*

Content moderation is outsourced to various companies in the Arab world. Vendors for tech companies around the world are paid millions of dollars annually to provide content moderation services to foreign tech companies (Satariano & Isaac, 2021). Study participants mentioned six different vendors located in four Arab countries in the region. These vendors, including Teleperformance and Concentrix in Tunisia and Teleperformance in Morocco, are largely responsible for the moderation of Arabic content for various tech companies. To get hired as a content moderator, an applicant needs only to demonstrate moderate proficiency in English-language skills and pass an interview with an HR representative and a psychologist. If they are hired, they undergo training that can last for about a month. In our interviews, participants noted that content moderators are instructed on various policies and are often required to memorize them during training. Then they have to pass tests demonstrating their familiarity with company policies and their understanding of the various dialects in the Arab world. If they pass both tests, they begin their content moderation job by evaluating and labeling the content they receive.

As in other regions, this job takes a severe toll on moderators (Dwoskin et al., 2019). **Many moderators we interviewed describe themselves as still recovering from the trauma they experienced. Participants reported having experienced anxiety, depression, and insomnia,**

with some requiring antidepressants to cope, even after resigning from their positions. In return for this taxing work, they receive a monthly salary equivalent to USD 400-600, with a potential USD 100 bonus based on reaching certain performance metrics. Participants regarded this compensation as low and inadequate for the trauma they experienced. All vendors have a psychology department to provide mental health support for moderators, although participants considered it largely ineffective. One moderator noted:

> *"It affected me and my mental health badly. Even psychologists didn't help. I reached the level that I had to take medications so I could get some sleep. It was that bad. Trust me, seeing this content once or twice is kinda normal. But when you see like 150 of this content every day for a year, it is not normal. It is not just that. It affects the way you think as well. You see people in another way because the people you see on the platform are way different from the people you know in real life." (Content Moderator, March 2024).*

Not only does the trauma make this job impossible to enjoy, moderators also have to deal with a strict evaluation system that is based on the number of videos they review per day, how long it takes them to make the decisions, and how accurate the decision is. Participants stated that they had to moderate hundreds of instances of disturbing content on a daily basis. A moderator at TikTok mentioned that they had to review about 170 videos per day, making a decision about each video within 20 seconds to a minute. The faster they completed the job, the better their metrics would look at the end of the week. In some instances, they had to speed up videos to make a decision faster. Despite the pressure to work quickly, any mistake could impact their monthly salary and might lead to termination.

B. A GLOBALIZED VS. LOCALIZED CONTENT MODERATION APPROACH

Through conversations with many content moderators and representatives from the T&S and policy teams at various tech companies, **we found that there are two approaches to content moderation: a) a global one, which entails employing the same policies on all users, and b) a localized one that tailors some of the policies for each region.** We found that TikTok generally adopts a more localized content moderation strategy, acknowledging the distinctiveness of each region and implementing tailored policies to accommodate regional differences and cultural nuances. According to participants, this is in an attempt to have a less restrictive content moderation approach to gain a strong audience base in Global South countries.

Conversely, Meta, X, and YouTube opt for a global content moderation approach, employing uniform policies on all users worldwide. This approach is easier to apply and aims towards equitable treatment of all content. A former employee at the policy team of one of these companies stated:

> *"Quite frankly, it's simpler for the company to have one version of a policy that they apply globally, and so making it the strictest version that applies to regulation in a major market, say, for instance, Europe, is much easier than having a completely separate policy for Europe that doesn't apply elsewhere." (Ex policy team member, April 2024).*

Other participants noted that global content moderation is how tech companies ensure that governments do not define the rules of content moderation, thereby helping to avoid human rights violations. While a global content moderation approach seeks to apply all policies to all countries with no distinction, these policies primarily adhere to laws from the Global North, where tech companies are held accountable. For example, Meta's contentious Designated Dangerous Individuals and Organizations (DIO) list, which has been accused of bias against Muslims and Arabs, follows the foreign policies of the US, UK, and EU (Biddle, 2021). A former Meta employee stated:

> *"The way we put together that list was basically by taking the US specially designated nationals list and the organizations that they designate, as well as the organizations that any specially designated nationals run and combining it with the EU list and the UK list. And so the bias you're seeing there is frankly a result of the bias in those lists." (Former Employee, Meta, April 2024).*

As mentioned earlier, TikTok applies a localized content moderation approach. TikTok Arabic content moderation divides the Middle East and North Africa (MENA) region into MENA 1 (Lebanon, Egypt, Palestine, Syria), MENA 2 (Sudan, Gulf, Yemen), and MENA3

(Tunisia, Morocco, Algeria, Mauritania, and Libya). Recently, MENA 1 and 2 were combined. A moderator at TikTok would be assigned to one or more sub-region, each of which has a slur list, policies, and cultural rules that are shared with the moderators. For example, a woman wearing short sleeves is unacceptable in MENA 2 but not in MENA 1 and 3. However, moderators expressed their confusion about these policies and their inability to understand all of these distinctive regional policies at once. Additionally, they felt that these policies lacked clear interpretations to the extent that some experienced moderators had to write an interpretation list to explain the policies to the newly hired. A moderator added:

> *"But so we were updating the [slur] spreadsheet. We were updating some of the organizations' names and flags and policies in general, like the understanding of the policies. They always had different interpretations. We were trying to keep up with their own interpretation of things." (Content moderator, April 2024)*

While TikTok applies a more localized approach, the other companies we examined in this report employ a more uniform approach to moderation, albeit with some variation by region/country. According to interviewees, Meta has a list of slur words for each country, which is updated frequently. Yet, there is no distinction between regions in relation to the employment of their policies.

Despite this, moderators at both TikTok and Meta reported that the variations in dialects across the Arab world often caused confusion. **Arabic-speaking moderators were assigned content from any country in the region, regardless of their native dialect, which frequently led to errors.** This approach assumed that all Arabs understand all Arabic dialects and contexts. One moderator noted:

> *"[Knowledge about Arabic dialects] wasn't, like, a criterion at the beginning, and you can't change it now. So you might as well just carry on with keeping a blind eye on it and just, you know, pretend that you're doing well. But, you know, ...most of the videos, most of the mistakes and the errors were done, that was done, they were due to the lack of [understanding of] the language." (Content moderator, April 2024).*

C. MAGHREBI ARABIC DIALECTS & AUTOMATED CONTENT MODERATION (ACM)

Automated tools are a necessary supplement to manual content moderation, given the scale of user-generated content around the world on social media (Shenkman et al., 2021). Automated content moderation (ACM) is the first layer of analyzing content. For example, Facebook uses a tool to assign predictive scores to posts, estimating their likelihood of violating terrorism policies. Posts with higher scores are prioritized for review by specialized human moderators (Gorwa et al., 2020). Although human reviewers usually make the final decision, ACM systems can automatically remove posts or reduce their visibility if they are highly confident that the content promotes terrorism (Gorwa et al., 2020).

If ACM cannot make a confident judgment, it will pass items to a moderator, who will further investigate it and make decisions. However, according to content moderators at TikTok, there is an exception for high-value users, typically those with hundreds of thousands of followers, whose content is exclusively reviewed by a human moderator.

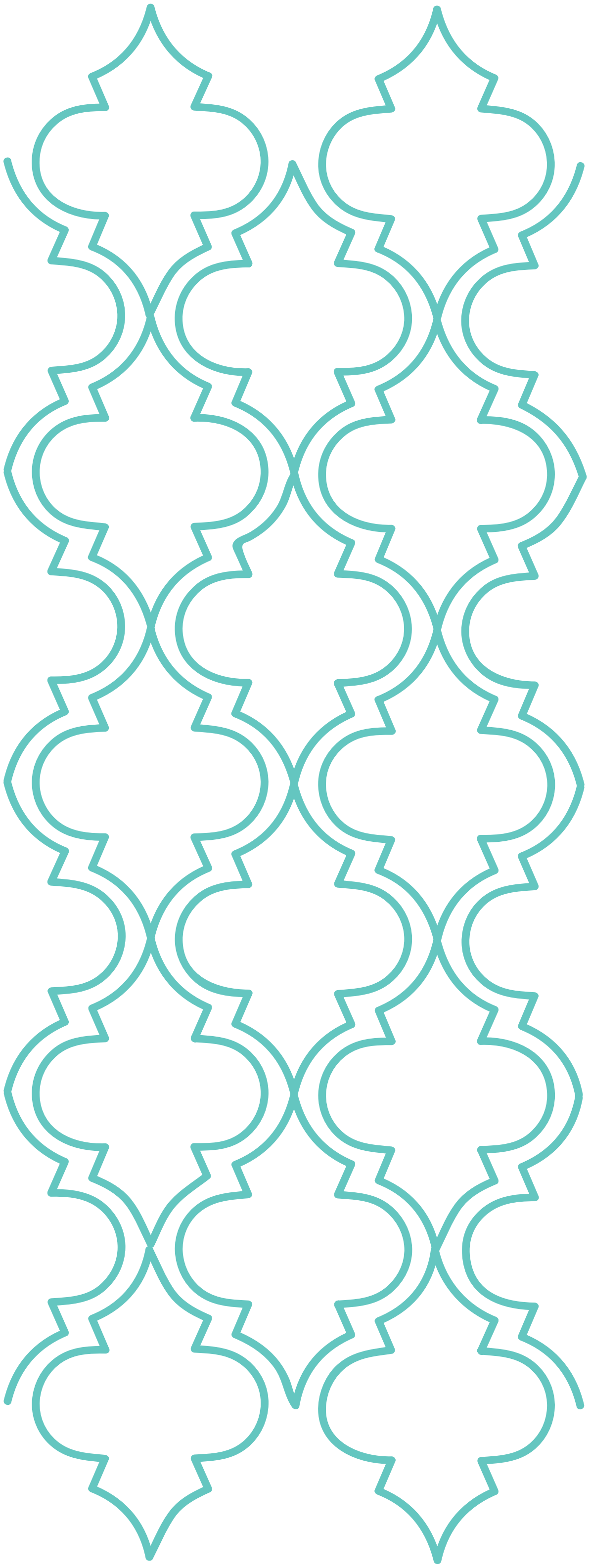
The Maghrebi Arabic dialects are low-resource ones. This "resourcedness" gap has led to models that do not perform well compared to Modern Standard Arabic (Fusha) or even other more popular dialects like the Egyptian dialect. Additionally, datasets lack sufficient examples of Arabizi and code-switching to train the classifiers. An NLP researcher who works at a social media company noted:

> *"Due to the lack of data, a lot of the models that are trained on Moroccan Arabic or even Algerian Arabic do not perform so well compared to the ones trained on Modern Standard Arabic."*
> *(NLP researcher, social media company, April 2024).*

Additionally, flawed data annotation, the process of labeling data to train AI models, can also lead to errors in automated content moderation (Binns et al., 2017). Several Maghrebi Arabic NLP researchers noted that the recruitment of annotators who do not speak Maghrebi dialects and the lack of funding and resources have led to inconsistencies, biases, and inaccuracies in the labeled datasets. These errors ultimately impact the quality of the classifiers and content moderation in general, resulting in incorrect decisions and unjustified content removals.

The lack of diversity in NLP research teams also harms the quality of classifiers. Former employees at different tech companies stated that there is no intentional diversity of linguistic backgrounds in the NLP teams and that language or cultural knowledge is not a requirement to get this job. While this is a common practice in the NLP community, it may lead to less rigorous classifiers for non-English low-resource languages, including Maghrebi Arabic dialects. For instance, many content creators in the Maghreb region noticed that Instagram instantly hides any comment that includes, "Allah Akbar" (translated as "God is the Greatest"), a sentence commonly used in the Arab world for prayer and to express happiness and joy, but which can also be associated with terrorist contexts. When asked about this phenomenon, one NLP researcher explained that this could be because “when you don't have someone in the room who can stop these things when the models are being trained... these are the results that you get.” (NLP researcher, social media company, April 2024).

Recommendations



It is critical that tech companies improve their moderation of Maghrebi Arabic dialects to avoid perpetuating the legacy of colonialism that has already marred the language, culture, and economy in this region. The errors in the algorithms and misunderstandings by moderators are issues that can be resolved with the right intentions and efforts, such as the following:

A. Recruit Native Annotators

Currently, tech companies rely on third-party companies for labeled data to train their AI models. These third-party companies often assign Arabic-speaking annotators from any country in the region to annotate any Arabic dialect. This practice leads to significant misinterpretation of the data and results in incorrect decisions that affect the accuracy of the AI training data. Annotators are typically better at identifying their own dialects and are more likely to confuse dialects with which they are unfamiliar (Abu Farha & Magdy, 2022). Native annotators will have a deeper understanding of cultural contexts, slurs, idioms, and other linguistically complex elements, which will improve the quality of labeled data used for training AI models. This includes understanding phenomena like Arabizi and code-switching. Hence, vendors should ensure the recruitment of native annotators with diverse demographic backgrounds to decrease the biases and errors in the data.

B. Prioritize Hiring a Diverse NLP Team

Tech companies should also prioritize hiring a diverse team of NLP scientists who reflect the various regions and cultures they serve, including the Maghreb region. By doing that, tech companies can ensure better model accuracy that is informed by a comprehensive understanding of cultural nuances. Additionally, tech companies need to benefit from the expertise of local researchers in the Maghreb region who have examined the unique characteristics of Maghrebi Arabic dialects, such as Arabizi and code-switching. This could be done through providing local research groups with technical and financial support.

C. Enhance Communication and Support Channels

One of the biggest concerns for users in this region is their inability to effectively communicate with tech companies when reporting violations, appealing decisions, or addressing issues like hacking. This lack of direct and efficient communication leaves users feeling vulnerable, particularly when dealing with urgent or serious matters. Tech companies should prioritize establishing robust communication channels through their regional offices to address these concerns. These communication channels need to be fast, responsive, and culturally sensitive to the unique needs of users in the region.

Additionally, the appeal and reporting processes should be more organized and clear to users. Users need to be able to track the status of their reports and appeals and understand what actions are being taken and whether it has been reviewed by the algorithm or by a moderator.

Also, tech companies should be more responsive to requests from civil society organizations that assist users on the ground and advocate on their behalf. These organizations play a crucial role in bridging the gap between users and tech companies, fostering a safer and more inclusive information environment in the Maghreb region.

D. Foster Transparency About Content Removal and Shadowbanning

Users in the Maghreb region frequently expressed frustration over unexplained content removals and shadowbanning, particularly with political posts. This uncertainty concerning the fate of their content has created an atmosphere of self-censorship, leading users to overthink and scrutinize their content before posting it. This hinders the free exchange of ideas, feelings, and emotions and contributes to a climate of online censorship.

Consequently, tech companies should prioritize transparency regarding content removals. This transparency should entail informing users when a moderation decision has occurred and the reason for the action. Transparency should include instances of demoting a user’s content. Companies should also clarify whether these restrictions were implemented by humans or algorithms. Such disclosure efforts will rebuild user trust in these services and replace the feelings of undue censorship.

Additionally, companies should only use shadowbanning in rare circumstances. It can be justified when necessary to prevent bad actors from misusing a service and is particularly useful in protecting users from spam, coordinated disinformation, and harassment by sockpuppet accounts. When employing shadowbanning, companies should disclose the specific situations that prompted its use.

E. Address Moderators’ Dialects and Expertise

Content moderators are overworked, underpaid, and exposed to disturbing content that leaves them traumatized. As a result, many moderators are forced out of the workforce to address the trauma they experience. The loss of trained moderators and their replacement with new hires adds to the challenges of content moderation since they take a lot of time to understand the policies and get used to the moderation systems. Content moderators report that they would have stayed if they were offered higher payments, better work-life balance, and better psychological support. Tech companies can reduce turnover by providing moderators with appropriate support. Having experienced moderators who can safely perform their jobs will inevitably improve the moderation system.

Additionally, while there are many vendors across the Arab world, moderators are often assigned to content outside their native dialect. This leads to numerous errors in moderators' decisions due to misunderstanding and lack of context. Instead, moderators should only make decisions on content in their own dialects.

Lastly, companies should provide better interpretation of policies in order for moderators to accurately enforce them. Training should include more regional examples with cultural contexts to help moderators understand the content they review and make accurate decisions.

Appendix

Methods and Data Collection

To study content moderation systems for the Maghrebi Arabic language, we used a mixed-method approach, combining qualitative and quantitative methods. First, we conducted semi-structured, in-depth interviews to assess the information ecosystem in the Maghreb region, the governance of Maghrebi Arabic by tech companies, and the challenges digital rights activists face in advocating for equitable moderation for the Maghrebi Arabic dialects. We were able to interview 14 participants from a variety of backgrounds. We interviewed six current and former Arabic content moderators from third-party vendors in the Maghreb region. We also interviewed six participants who worked or are still working in the policy, Trust & Safety, or research teams at Meta, Twitter, YouTube, and TikTok. Lastly, we interviewed two leading digital rights advocates from the Maghreb region.

All these interviews were conducted online. We referred to all participants in the study using pseudonyms. The interviews occurred between March and May, 2024. The interviews were predominantly conducted in English, with one conducted in Arabic. Field notes were taken during the interviews, and the interview recordings were later transcribed and analyzed to find common themes among the participants.

In addition to these interviews, we conducted five focus group sessions that included internet users from Tunisia, Morocco, and Algeria. With the help of our regional partner, “Digital Citizenship” in Tunisia, we were able to sit down and speak with more than 25 participants from the Maghreb region to discuss their day-to-day uses of online services, the obstacles they face, their knowledge of the community standards, and how they evade the algorithms. These focus groups were held online, and each session lasted approximately two hours.

Lastly, we conducted an online survey to understand how Internet users in the Maghreb region use online services, report inappropriate content, and deal with content removals. The survey asked questions about users’ trust in social media companies and their perceptions of the toxicity of online platforms in their region. The Alchemer platform was used to distribute the online survey from April 1st to May 5th, 2024. A modest honorarium of US$10 was offered for participating. The survey was administered in English, Arabic, and French. We were able to collect a sample from 111 participants from the Maghreb region. About 47% of the participants were from Tunisia, 30% from Morocco, and 23% from Algeria. The majority of the survey participants self-identified as females from 25 to 35 years old.

We have also organized a roundtable discussion with Natural Language Processing (NLP) researchers from the Maghreb region and the broader Arab world to address the challenges they face in data collection, annotation, and processing.

References

Abu Farha, I., & Magdy, W. (2022). The Effect of Arabic Dialect Familiarity on Data Annotation. In H. Bouamor, H. Al-Khalifa, K. Darwish, O. Rambow, F. Bougares, A. Abdelali, N. Tomeh, S. Khalifa, & W. Zaghouani (Eds.), *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)* (pp. 399–408). Association for Computational Linguistics. [perma.cc/V94X-FMEF]

Biddle, S. (2021, October 12). Revealed: Facebook's Secret Blacklist of "Dangerous Individuals and Organizations." The Intercept. [perma.cc/8JG3-3UQ5]

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017). Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. In G. L. Ciampaglia, A. Mashhadi, & T. Yasseri (Eds.), *Social Informatics* (pp. 405–415). Springer International Publishing. [perma.cc/AP5E-5BQT]

Darwish, K. (2014). Arabizi Detection and Conversion to Arabic. In N. Habash & S. Vogel (Eds.), *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)* (pp. 217–224). Association for Computational Linguistics. [perma.cc/6FSR-LX9J]

Duwairi, R. M., Alfaqeh, M., Wardat, M., & Alrabadi, A. (2016). Sentiment analysis for Arabizi text. *2016 7th International Conference on Information and Communication Systems (ICICS)*, 127–132. [perma.cc/5BNK-J2QR]

Dwoskin, E., Whalen, J., & Cabato, R. (2019, July 25). Content moderators at YouTube, Facebook and Twitter see the worst of the web and suffer silently. *Washington Post.* [perma.cc/8A3G-X72U]

Elswah, M. (2023). *Online tactical innovation and stagnation: Insights from the aftermath of the Arab Spring in Syria and Tunisia* [Ph.D., University of Oxford]. [perma.cc/VY3Y-5ZJ6]

Elswah, M. (2024, January 30). Investigating Content Moderation Systems in the Global South. *Center for Democracy and Technology.* [perma.cc/8DUF-ZTJF]

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society, 7(1).* [perma.cc/Z6YG-3GZE]

Grimmelmann, J. (2015). The Virtues of Moderation. *Yale Law & Tech,* 17, 42–109. [perma.cc/39TM-4BV4]

Guellil, I., Saâdane, H., Azouaou, F., Gueni, B., & Nouvel, D. (2021). Arabic natural language processing: An overview. *Journal of King Saud University - Computer and Information Sciences, 33(5)*, 497–507. [perma.cc/7ABJ-RPN2]

Hajbi, S., Chihab, Y., Ed-Dali, R., & Korchiyne, R. (2022). Natural Language Processing Based Approach to Overcome Arabizi and Code Switching in Social Media Moroccan Dialect. In Y. Maleh, M. Alazab, N. Gherabi, L. Tawalbeh, & A. A. Abd El-Latif (Eds.), *Advances in Information, Communication and Cybersecurity* (pp. 57–66). Springer International Publishing. [perma.cc/362K-WGJP]

Harrat, S., Meftouh, K., & Smaïli, K. (2018). Maghrebi Arabic dialect processing: An overview. *Journal of International Science and General Applications, 1.* [perma.cc/6KNS-EURE]

Lorenz, T. (2022, April 11). Internet 'algospeak' is changing our language in real time, from 'nip nops' to 'le dollar bean.' *Washington Post.* [perma.cc/8YZW-ZNWB]

Lorenz, T., & Nix, N. (2024, February 11). Meta turns its back on politics again, angering some news creators. *Washington Post.* [perma.cc/67K8-PG8J]

Nicholas, G. (2022). Shedding Light on Shadowbanning. Center for Democracy & Technology. [perma.cc/577H-P23T]

Nicholas, G., & Bhatia, A. (2023). *Lost in Translation: Large Language Models in Non-English Content Analysis.* Center for Democracy & Technology. [perma.cc/7LLC-5Z5S]

Roberts, S. T. (2018). Digital detritus: "Error" and the logic of opacity in social media content moderation. *First Monday.* [perma.cc/A7VB-4Q93]

Satariano, A., & Isaac, M. (2021, August 31). The Silent Partner Cleaning Up Facebook for $500 Million a Year. *The New York Times.* [perma.cc/WLY6-CMV4]

Shenkman, C., Thakur, D., & Llansó, E. (2021). *Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis.* Center for Democracy & Technology. [perma.cc/9DY6-8324]

Younes, J., Souissi, E., Achour, H., & Ferchichi, A. (2020). Language resources for Maghrebi Arabic dialects' NLP: A survey. *Language Resources and Evaluation,* 54(4), 1079–1142. [perma.cc/V7W9-5F9D]

cdt.org

cdt.org/contact

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

202-637-9800

@CenDemTech

CENTER FOR
DEMOCRACY
& TECHNOLOGY