

Hate Speech in Lebanon



**The Shortcomings
and Responsibilities
of Social Media Platforms**



July 2023



THE SAMIR KASSIR FOUNDATION

In partnership with:



Acknowledgements

Afef Abrougui, the owner of Fair Tech, a consultancy based in the Netherlands with the mission to protect human rights in the digital space, led this research and compiled and edited the report.

Mirna Ghanem, Senior Researcher at Samir Kassir Foundation, conducted step 1 research, collected hate speech cases, and conducted Lebanon research.

Farah Rasmi, a PhD candidate in International Relations/Political Science at The Graduate Institute of International and Development Studies, Geneva, conducted step 1 and step 2 research, in addition to the comparative analysis.

The Samir Kassir Foundation (SKF) is a non-profit civic organization, working within the civil society and cultural circles to spread the democratic culture in Lebanon and the Arab world, through monitoring and reporting on violations of freedom of expression in Lebanon and the Arab region, empowering journalists and emerging digital media outlets, and involving media outlets, civil society organizations, and policy makers in the debate on media policy reforms.

Ranking Digital Rights (RDR) is a non-profit which produces a Corporate Accountability Index ranking the world's most powerful internet, mobile ecosystem, and telecommunications companies on relevant commitments and policies, based on international human rights standards.

Table of Contents

About this Study	5
Key Findings	5
Methodology	8
Introduction	12
Discrimination and Hate Speech in Lebanon	15
The Availability of Key Policies in Arabic	19
Indicator-based Findings: Where Each Platform Stands	24
How Platforms Define Hate Speech and Enforce Rules	32
How Content Is Identified and Removed	33
Partnerships with Civil Society Organizations	34
Recommendations for Platforms	37
Annexes	39
Appendix 1: List of Key Terms and Definitions	39
Appendix 2: Research Indicators and Elements	41



About this Study

This research project evaluates the hate speech policies of four social media platforms: Facebook, Twitter, YouTube, and TikTok, as implemented in Lebanon, using a selection of Ranking Digital Rights’ human rights-based indicators. Facebook, Twitter, and YouTube are owned respectively by U.S. technology companies Meta Platforms Inc. (known as Meta), X Corp., and Google LLC, a subsidiary of Alphabet Inc. TikTok is owned by the Chinese technology company ByteDance.

RDR’s methodology is employed to benchmark companies in the ICT sector using indicators that establish high yet attainable standards for corporate transparency and policies that align with internationally recognized human rights standards.

The research focuses on hate speech policies applied to both user-generated and advertising content, shedding light on the policies’ potential effectiveness, transparency, user-friendliness, fairness, and respect for freedom of expression and the right to non-discrimination. We also sought to document any significant disparities in policies’ availability in Arabic and English.

We selected Facebook, TikTok, and YouTube for their widespread usage in Lebanon, with a combined user base nearing 10 million as of [early 2023](#). In 2022, Lebanon had [a population](#) of 6.7 million people. Although Twitter has significantly fewer users in Lebanon – 531 thousand users – we decided to include it for two reasons. Firstly, in Lebanon and elsewhere in the region, Twitter serves as a platform for [political debate and mobilization](#). Secondly, following Elon Musk’s takeover, we were interested in examining any potential changes in the company’s hate speech policies and their implications for a deeply divided country like Lebanon.

Platform	Company	Home market	Number of users in Lebanon as of early 2023 (Source datareportal.com)
Facebook	Meta Platforms, Inc.	United States	2.95 million users
TikTok	ByteDance Ltd.	China	2.78 million users
Twitter	X Corp.	United States	531 thousand users
YouTube	Google LLC	United States	4.91 million users

Table 1

Key Findings

Access to Policies in Arabic

- In general, all four platforms have the majority of their available policies translated into Arabic. When an Arabic version of the policy is available, it is typically a direct translation into classical Arabic with little to no difference from the English version. The platforms make the Arabic language policies accessible (when available) through a simple language switch in the page settings. However, in some cases, key policies are not accessible in Arabic, such as TikTok's Intellectual Property Policy and Google's AI principles.
- Twitter's Terms of Service, which govern users' access to and use of the service, are not provided in Arabic. This creates a barrier for users in Lebanon (and elsewhere) who are only fluent in Arabic, as they cannot give informed consent when signing up for the service.
- Among the platforms, Facebook exhibits the most inconsistencies between its Arabic and English policies. In six out of 19 indicators, it provides less to no information at all in Arabic compared to its performance on these indicators when applied to its English-language policies.

Human Rights Practices

- Among the platforms studied, TikTok was the only one that did not explicitly and clearly commit to upholding human rights. Its policy did not encompass freedom of expression and information, and although it expressed a commitment to protecting the right to privacy, this commitment was not grounded in international human rights standards.
- There was no evidence to suggest that any of the companies owning the platforms included in the study conduct due diligence in Lebanon. None of the platforms conduct robust human rights impact assessments to understand how their policy enforcement processes affect the fundamental rights of their users in Lebanon, particularly communities at higher risk of experiencing hate speech, such as migrant workers, refugees, LGBTQ+ community, and women. Consequently, they failed to address and mitigate the negative impacts that arise from these risks.

Twitter under Musk

- Under the leadership of Elon Musk, Twitter has experienced setbacks in terms of freedom of expression and protection from hate speech. Since 2021, the company has ceased publishing its transparency reports on Rules Enforcement and Removal Requests. These reports provide insights into the actions taken by the company to restrict content and enforce its rules, as well as its response to third-party demands. Furthermore, Twitter disbanded its Trust and Safety Council, which previously brought together civil society representatives from various regions worldwide, including a Lebanese NGO, to provide advice on the platform's rule development and product enhancements.

- The implications of Musk’s takeover and the changes he implemented on the spread of hate speech in Lebanon remain unclear. However, with the discontinuation of transparency reports on content moderation actions, it has become increasingly challenging for researchers and civil society to monitor how the platform handles hate speech.

Algorithmic Transparency

- All of the platforms included in the study utilize machine learning algorithms for various purposes, including content ranking and moderation. However, despite the human rights risks associated with these systems, none of the companies explicitly and clearly articulated a policy commitment to human rights in the development and utilization of their algorithmic systems. While Meta and YouTube provided commitments that were not clearly grounded in human rights principles, TikTok and Twitter did not make any commitments at all.
- Platforms lacked transparency regarding their use of algorithms to curate, recommend, and rank content. While TikTok disclosed more details compared to its counterparts, including the variables that influence ranking systems and user options to control those variables, this information was not available in Arabic. Facebook only provided information about how its Feed curates and recommends content using algorithmic systems, without specifying how it uses these systems in other areas such as friend recommendations and search results.
- Platforms exhibited even less transparency concerning their policies governing the use of bots. Twitter was the most transparent, disclosing clear rules and enforcement mechanisms. TikTok and Meta disclosed almost no information, while YouTube did not disclose anything regarding their bot policies.

Targeted Advertising

- Among all the platforms in the study, Twitter and YouTube demonstrated the highest level of transparency regarding their ad content and ad targeting policies.
- TikTok and YouTube were the only platforms that published data about the volume and nature of actions taken to restrict advertising that violated their policies. However, the data they provided was not comprehensive and did not disaggregate advertisements rejected for violating ad content rules from those rejected for violating ad targeting rules.

Censorship Requests

- Platforms are not explaining the processes they follow to handle content restriction requests pertaining to hate speech. We are aware that technology platforms, including Facebook, TikTok, Twitter, and YouTube, have partnerships with civil society organizations under the 2016 “EU Code of conduct on countering illegal hate speech online.” These partnerships involve organizations submitting reports of hateful content. However, it remains unclear how the platforms assess these requests before responding. It is also unclear whether the platforms receive requests from private sector entities in Lebanon to restrict hateful content.

- All platforms publish data on government demands they receive to restrict content and accounts, with YouTube providing the most comprehensive data, including information on the types of subject matter associated with these demands. Facebook, TikTok, and Twitter do not specify the subject matter, making it challenging for researchers, advocates, civil society, and journalists in Lebanon and elsewhere to understand the extent to which these demands are related with hate speech.

Methodology

The research methodology employed in this study is an adaptation of RDR's methodology, which consists of [58 human rights-based indicators](#) organized under three main categories: Governance, Freedom of Expression and Information, and Privacy. These indicators aim to assess whether companies uphold the rights to freedom of expression, information, and privacy, as outlined in the Universal Declaration on Human Rights (UDHR) and the International Covenant on Civil and Political Rights (ICCPR).

For the purpose of this research, we focused on 19 specific indicators to gain a deeper understanding of the extent to which users in Lebanon are kept informed about policies that affect their exposure to hate speech on Facebook, Twitter, YouTube, and TikTok, as well as how these platforms moderate and censor such content. The relevant indicator areas and indicators are listed below, and a comprehensive list of the research indicators and their elements can be found in Appendix 2. Appendix 1 contains a list of key terms and definitions.

Governance

- **G1. Policy Commitment.** The company should publish a formal policy commitment to respect users' human rights to freedom of expression and information and privacy.
- **G4(b). Impact assessment: Processes for policy enforcement.** The company should conduct regular, comprehensive, and credible due diligence, such as through robust human rights impact assessments, to identify how its processes for policy enforcement affect users' fundamental rights to freedom of expression and information, to privacy, and to non-discrimination, and to mitigate any risks posed by those impacts.
- **G6(b). Process for content moderation appeals.** The company should offer users clear and predictable appeals mechanisms and processes for appealing content-moderation actions.

Freedom of Expression and Information

- **F1(a, b, c): Access to policies.** Companies should ensure that their policies affecting users' freedom of expression and information are easy to find and easy to understand. We expect companies to make their terms of service policies (Indicator F1a), advertising content policies (Indicator F1b), and advertising targeting policies (Indicator F1c) easy to access, available in Arabic, and presented in an understandable manner.

- **F3(a). Process for terms of service enforcement.** The company should clearly disclose the circumstances under which it may restrict content or user accounts.
- **F3(b). Advertising content rules and enforcement.** The company should clearly disclose its policies governing what types of advertising content is prohibited.
- **F3(c). Advertising targeting rules and enforcement.** The company should clearly disclose its policies governing what type of advertising targeting is prohibited.
- **F4(a). Data about content restrictions to enforce terms of service.** The company should clearly disclose and regularly publish data about the volume and nature of actions taken to restrict content that violate the company’s rules.
- **F4(b). Data about account restrictions to enforce terms of service.** The company should clearly disclose and regularly publish data about the volume and nature of actions taken to restrict accounts that violate the company’s rules.
- **F4(c). Data about advertising content and advertising targeting policy enforcement.** The company should clearly disclose and regularly publish data about the volume and nature of actions taken to restrict advertising content that violate the company’s advertising content policies and advertising targeting policies.
- **F5(a). Process for responding to government demands to restrict content or accounts.** The company should clearly disclose its process for responding to government demands (including judicial orders) to remove, filter, or restrict content or accounts.
- **F5(b). Process for responding to private requests for content or account restriction.** The company should clearly disclose its process for responding to requests to remove, filter, or restrict content or accounts that come through private processes.
- **F6. Data about government demands to restrict content and accounts.** The company should regularly publish data about government demands (including judicial orders) to remove, filter, or restrict content and accounts.
- **F7. Data about private requests for content or account restriction.** The company should regularly publish data about requests to remove, filter, or restrict access to content or accounts that come through private processes.
- **F12. Algorithmic content curation, recommendation, and/or ranking systems.** Companies should clearly disclose how users’ online content is curated, ranked, or recommended.
- **F13. Automated software agents (“bots”).** Companies should clearly disclose policies governing the use of automated software agents (“bots”) on their platforms, products and services, and how they enforce such policies.

Privacy

- **P1(b). Access to algorithmic system development policies.** The company should offer algorithmic system development policies that are easy to find and easy to understand.

Research Threads

The project methodology follows three threads: indicator-based research, comparative content analysis, and report writing.

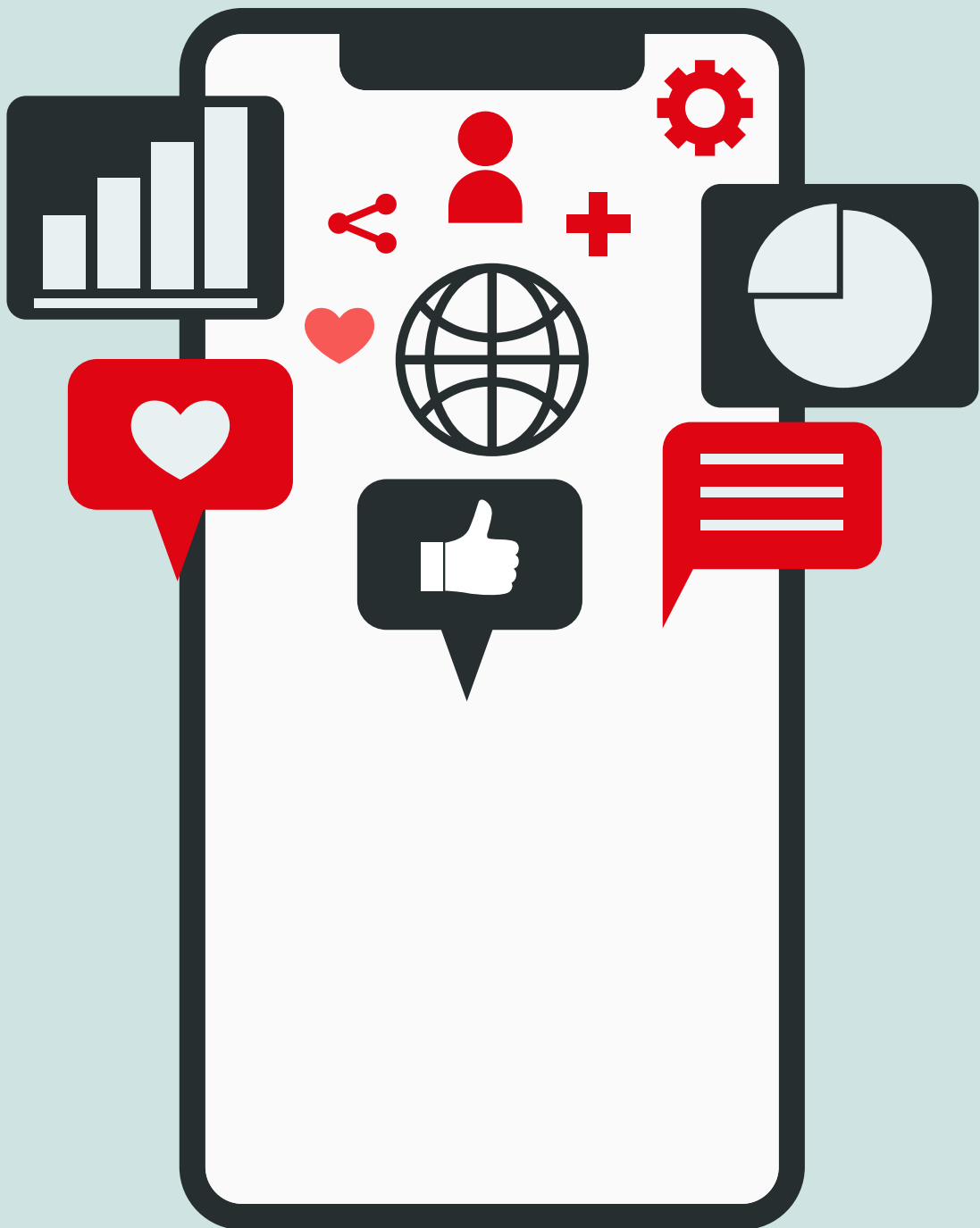
For the indicator-based research, we utilized the aforementioned indicators to assess the policies of each platform in both Arabic and English. This allowed us to identify any significant differences in the Arabic policies, considering that Arabic is the official and primary language spoken in Lebanon. We considered documents and policies published from December 12, 2019 to December 12, 2022. The indicator-based research involved the three following steps:

- Step 0: A researcher copied the latest data previously collected by RDR from a separate past research [project](#) into an input sheet.
- Step 1: A primary researcher verified the accuracy of the data from Step 0. If accurate, it was copied into Step 1. If not, a new assessment was written.
- Step 2: A different researcher, who was not involved in Step 1 for that specific indicator and company, meticulously fact-checked the data from Step 1. If the Step 2 researcher agreed with the analysis, it was copied into Step 2. If there was disagreement, they engaged in discussions with the Step 1 researcher until reaching a decision, which was then recorded in Step 2.

Following the indicator-based research, we conducted comparative analysis to:

1. Identify any language variations in policy disclosures for each platform.
2. Gather the relevant policies of each platform pertaining to hate speech, including those related to content moderation of hate speech and the use of Artificial Intelligence (AI) in such moderation.
3. Identify discrepancies in the actual content of the platforms' policies regarding hate speech, including how hate speech is defined and how the enforcement of hate speech, the role of content moderation, and the use of AI in such moderation are described. We also assessed whether and how they address hate speech harms specifically relevant to Lebanon.
4. Identify disparities in the platforms' disclosures, including which platform provided the most comprehensive disclosure and which platforms disclose the least in key areas related to hate speech and freedom of expression.

Finally, during the report writing phase, we utilized the results of the indicator-based research, comparative analysis, and desk research on human rights and hate speech online in Lebanon to compile our findings.



Introduction

International human rights law does not explicitly define hate speech, but rather addresses various forms of discrimination and focuses on severe aspects of hate speech, such as incitement to violence and propaganda. The International Convention on the Elimination of All Forms of Racial Discrimination [defines](#) racial discrimination as “any distinction, exclusion, restriction or preference based on race, colour, descent, or national or ethnic origin which has the purpose or effect of nullifying or impairing the recognition, enjoyment or exercise, on an equal footing, of human rights and fundamental freedoms in the political, economic, social, cultural or any other field of public life.”

Article 4 of the Convention [prohibits](#) “all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination,” and incitement to violence “against any race or group of persons of another colour or ethnic origin.” It also prohibits propaganda activities that “promote and incite racial discrimination.”

International human rights instruments also seek to strike a balance between individuals’ rights to express themselves and their responsibility not to harm others through discriminatory statements and speech that constitute incitement. The [International Covenant on Civil and Political Rights](#) (ICCPR) enshrines freedoms of opinion, expression, and information under Article 19, while prohibiting “any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence.”

Hate speech, however, is a broad term that [encompasses](#) any speech that negatively singles out individuals or groups based on one or several aspects of their identity. This includes stereotypical comments targeting members of a certain faith, sexist and misogynistic speech targeting women, and other dangerous forms of hate speech, such as incitement to violence against individuals based on their sexual orientation or gender appearance, or calls for mass murder against a minority group.

Even when hate speech does not involve incitement to acts of violence, it can still have harmful effects on victims and social cohesion. Online hate speech, in particular, is [harmful](#) to the mental health and well-being of those subjected to it. It further reinforces the exclusion and marginalization of already disadvantaged groups and communities, hindering their ability to fully exercise and enjoy their fundamental rights.

Hate speech in Lebanon disproportionately impacts women, refugees, migrant workers, and LGBTQ+ communities. Perpetrators of hate speech, both online and offline, are almost never held accountable, while individuals expressing themselves online may face hateful comments, attacks, and possibly legal repercussions, especially when they cross red lines and criticize those in positions of power.

While Lebanon criminalizes defamation and insults under its laws, discrimination and hate speech are not adequately addressed. [Vague defamation provisions in the Penal Code](#) can lead to imprisonment for up to three years. However, these provisions are frequently [exploited](#) by politicians and government officials to suppress criticism and silence those who expose corruption and the government’s failure to address the country’s and its population’s suffering.

In this context, social media platforms have a responsibility to balance the protection of users from hate speech while upholding their rights to freedom of expression and information. However, platforms are currently falling short of providing sufficient protections for users in Lebanon (and beyond), as this study will demonstrate.

A [report](#) by Ranking Digital Rights titled “It’s Not Just the Content, It’s the Business Model: Democracy’s Online Speech Challenge” highlights a crucial aspect of today’s global challenge posed by online content: “As a society, we are facing a problem stemming not just from the existence of disinformation and violent or hateful speech on social media, but from the systems that make such speech spread to so many people. We know that when a piece of content goes viral, it may not be propelled by genuine user interest alone. Virality is often driven by corporate algorithms designed to prioritize views or clicks, in order to raise the visibility of content that appears to inspire user interest. Similarly, when a targeted ad reaches a certain voter, and influences how or whether they vote, it is rarely accidental. It is the result of sophisticated systems that can target very specific demographics of voters in a particular place.”

Platforms build their algorithms in this way to generate profit. By driving engagement and keeping users actively involved through clicking, commenting, and posting, platforms harvest valuable data on user behavior, interests, preferences, wants and want-nots. Platforms then use such troves of data to tailor targeted advertisements to users, increasing the effectiveness of ad campaigns.

The report begins with an overview of discrimination and hate speech in Lebanon, focusing on the communities most affected. Chapter 3 maps the availability of platform policies governing hate speech in Arabic, documenting any discrepancies compared to the English-language policies. In Chapter 4, we analyze the platforms’ hate speech policies using the RDR indicators, assessing their application to user-generated content and advertising content. Chapter 5 delves deeper into the enforcement of these hate speech policies. The report concludes with a set of recommendations for the platforms.



Discrimination and Hate Speech in Lebanon

Despite Lebanon's ratification of the ICCPR and the International Convention on the Elimination of All Forms of Racial Discrimination, the country's legislation does not recognize nor adequately address discrimination and hate crimes. Instead, it relies on broad provisions that prohibit actions and statements that incite sectarian or racial strife.

Article 317 of the [Lebanese Penal Code](#) criminalizes “every action, every writing, and every speech intended or resulting in incitement to sectarian or racial strife or incitement to conflict between sects and the various elements of the nation.” Perpetrators can be sentenced to imprisonment for one to three years and fined between one hundred to eight hundred thousand Lebanese pounds. While this provision is broad and can be exploited to restrict legitimate speech, it fails, like other Lebanese laws, to define and address discrimination based on race, nationality, sexual orientation, ethnicity, migration status, and other attributes.

Lebanon has faced a series of crises since 2019, including political turmoil after the October 2019 uprising, the COVID-19 pandemic, the [Beirut port explosion](#), and an [economic crisis](#) classified by the World Bank as one of the worst since the 19th century. These crises have further [marginalized](#) women and disenfranchised groups. In addition to facing violence and discrimination offline, they are also subjected to hate speech, including on social media platforms.

During the COVID-19 outbreak, women experienced increased violence as they were confined at home with their abusers. The economic crisis has disproportionately affected women, who face further discrimination in the job market. Women politicians and media professionals are also targeted with hate speech campaigns that utilize derogatory and misogynistic language. Over a two-year period from the end of 2020 to 2022, the Samir Kassir Foundation (SKF) monitored hate speech against marginalized groups on traditional media, Twitter, and Facebook. The study used specific keywords associated with each group. The [findings](#) revealed that supporters of traditional parties, particularly Hezbollah, a Shia Islamist political party and militant group, and the Free Patriotic Movement (FPM), a nationalist Christian-majority party founded by former President Michel Aoun, targeted outspoken women who criticized them. These attacks involved misogynistic slurs and insults aimed at undermining prominent women's opinions and damaging their credibility. Journalist Dima Sadek, for example, frequently became a victim of misogynistic language, with derogatory terms like “[femoids](#)” used to belittle her as a woman. These insults were often accompanied by inflammatory rhetoric falsely accusing her of being an “agent” of foreign countries whenever she criticized Hezbollah.

The refugee crisis is being [exploited](#) by the Lebanese authorities, which scapegoat refugees for the economic crisis as a means of diverting attention from the government's own responsibility. This discourse has translated into hate speech campaigns within Lebanese society, with discussions revolving around the repatriation of refugees to Syria as a solution to the economic crisis. As a result, refugees face daily violence, threats, and xenophobia. Some municipalities and Lebanese residents have expelled them from refugee camps and denied them their basic rights.

Despite pleas from UN agencies and human rights organizations urging against the repatriation of refugees to Syria, the Lebanese government put out a plan for repatriation and is cooperating with the Syrian government to implement it, claiming that Syria is “safe” for their return. UN calls to protect the refugees have been met with hate speech campaigns and negative comments on social media. For example, in response to a Facebook post by pro-FPM TV station OTV, on August 18, 2022, readers accused UN agencies of being politicized and alleged a hidden agenda to keep Syrian refugees in Lebanon. One user even called for setting refugee camps on fire to force the refugees to return to Syria. On Twitter, a campaign under the hashtag “#ارضنا_مش_للنازح_السوري” (our land is not for the displaced Syrian) targeted refugees in July 2022. Another user spread unsubstantiated information, claiming that if 10% of the two million refugees in Lebanon were trained in using weapons, it would lead to another armed conflict against the Lebanese population.

According to data collected by SKF between 2020 and 2022, refugees were the most targeted group for hate speech in both traditional media and social media, followed by the LGBTQ+ community. During the May 2022 election period, hate speech against refugees increased, particularly on social media, originating from two rival Christian right-wing political parties, the Lebanese Forces and FPM. However, hate speech campaigns against refugees reached a peak in August 2022 when the government-initiated negotiations with the Syrian government for their immediate repatriation.

Migrant workers in Lebanon endure harsh working conditions and are [deprived of their rights](#) under the *Kafala* system. This [system](#), often likened to modern slavery, exploits migrant domestic workers and prevents them from leaving their jobs without their employers’ consent. As a result, domestic workers are subjected to long working hours, denied holidays, and withheld salaries. In some cases, these conditions have driven victims to commit or attempt suicide. The Lebanese economic crisis that began in 2019 [has further exacerbated their situation](#), as many middle-class families had to let go of their domestic workers due to their inability to afford their wages in US dollars. Moreover, heads of households have withheld wages owed for previous months and years. Hate speech and incitement against migrant workers have become pressing issues, fueled by pervasive societal and institutional racism. The lack of media coverage of their rights and struggles contributes to further marginalizing this group. For instance, a Facebook [user](#) commented on a post about domestic migrant workers: “I don’t understand how people can drink a cup of tea prepared by them [migrant workers],” before suggesting hiring someone Lebanese because they are “clean,” unlike migrant workers who “disgust” the commenter.

Furthermore, SKF’s [study](#) also highlights the complete disregard for LGBTQ+ issues on all local television channels. This omission further marginalizes the LGBTQ+ community. On Facebook, the monitored content revealed that although the community lacks visibility on the platform, the latter serves as a fertile ground for incitement and hate speech. During the 2022 election period, independent candidates who supported LGBTQ+ rights faced attacks from supporters of traditional political parties. Additionally, LGBTQ+-related language, such as the word “gay,” is used as an insult and negative epithet against opponents in political debates on social media platforms.

Finally, issues concerning [people with disabilities](#) are generally overlooked in both traditional and social media. They receive limited coverage, mainly on specific occasions such as the International Day of Persons with Disabilities. Although this marginalized group does not experience the brunt of hate speech as other groups like refugees, LGBTQ+ communities, and migrant workers on social media, terms such as “handicapped” are used pejoratively to undermine others’ opinions, particularly in political debates.



A-Z


A
B
C

The Availability of Key Policies in Arabic: Documenting Inconsistencies and Major Differences

Overall, all four platforms had the majority of their available policies translated into Arabic. Where the Arabic language policy was available, it was translated into classical Arabic with little to no difference from the English language version of the policy. Users could easily access the Arabic policies, when available, by switching the language settings on the platforms. However, there were instances where key policies and reports were not available in Arabic. Most notably, [Twitter's Terms of Service](#) were not available in Arabic, posing a challenge for users in Lebanon and in the region who are not fluent in other languages to give informed consent when signing up. Additionally, a number of transparency reports, which provide crucial data on the nature and volume of actions that platforms take to restrict content and accounts, including data on [enforcement of advertising rules](#) (YouTube) and data on the number of [government](#) and [private requests](#) to restrict content and accounts (Facebook) were not available in Arabic. This lack of data hampers the understanding of the scope of censorship practices and compliance with third-party removal requests for important stakeholders in Lebanon, including journalists and activists.


Below, we provide a breakdown of the main differences in policy availability per platform.

Facebook



Facebook exhibited the highest number of inconsistencies between its Arabic and English policies. Across six out of 19 indicators, it either provided less information or completely omitted it in Arabic, in contrast to its English-language policies. These inconsistencies primarily stem from Meta, Facebook's parent company, not publishing crucial reports that provide data on the enforcement of its rules requests, both from private entities and governments, to restrict content. Specifically, the [Meta Human Rights Report](#), [Community Standard Enforcement Report](#), and [Intellectual Property Report](#) were not available in Arabic. The absence of the [Intellectual Property \(IP\) Transparency Report](#) in Arabic meant that Meta's policy regarding handling IP requests, including commitments to diligence on these requests and pushback against overboard ones, was not accessible to Arabic-speaking users. Additionally, clear guidance or examples pertaining to Meta's process of responding to government demands were not provided in Arabic. Furthermore, the [Content Restrictions Report's case studies page](#), which offered examples of handling such requests, was exclusively available in English.

TikTok



TikTok has a majority of its policies translated into Arabic. In cases where the policy was not available in Arabic, other documents offering the same information were provided, enabling the company to earn similar scores in specific areas, such as its human rights commitment (G1) and its mechanism allowing users to access remedy for freedom of expression related grievances (G6b).

However, TikTok's [Intellectual Property Policy](#) was not available in Arabic. Moreover, the platform provided less information in Arabic regarding its use of algorithmic systems. Specifically, it failed to provide information in Arabic on how these systems are employed to curate and recommend content (F12). This information is crucial for users, media, and civil society to understand the factors determining the content they see on the platform. Additionally, TikTok did not provide information in Arabic on how automation is utilized to detect and remove Child Sexual Abuse Material (F3a).


Twitter



Twitter generally had most of its policies translated into Arabic. However, there were some notable gaps. The [Terms of Service](#) were unavailable in Arabic. The page “[Twitter for Good](#),” where the company states its commitment to freedom of expression and human rights (G1), was also not available in Arabic either. Nevertheless, Twitter does make a commitment to human rights in a separate document titled “[Defending and respecting the rights of people using our service](#),” which is available in Arabic.

Furthermore, Twitter did not provide Arabic translations for its [Ads Transparency report](#) (F4c) and its [Global Impact Report](#). The latter included information about Twitter's use of algorithmic systems in its content moderation practices (F3a). Little information was also available in Arabic regarding the company's enforcement of its bot policies (F13).

YouTube



Almost all policies were available in Arabic for YouTube. However, [Google's AI Principles](#) document, which outlines the mother company's “commitment to develop technology responsibly and establish specific application areas [Google] will not pursue,” is not available in Arabic (G1, E3). Additionally, YouTube did not provide a translation of its “[Our Annual Ads Safety Report](#),” which contains data on actions taken to enforce its ad policies.

One notable issue with YouTube's Arabic-language policies is related to accessibility. The terms of services (ToS) are only available in Arabic if users change their entire language settings to Arabic, and there is no language switch option on the actual ToS page. Furthermore, it was discovered through further research that accessibility to the Arabic language policy is also determined by geo-location. In other words, even if the language settings are changed to Arabic, if the location is not set to a MENA region, the ToS will appear in the main language of the region or country, not in Arabic. This means that Arabic speakers outside of the MENA region would not have access to the Arabic ToS unless they change their location.

The table below summarizes the key policies that were not made available by platforms in Arabic.

Platform	Key policy not available in Arabic	Description of the policy or document
Facebook	Meta's human rights report	The first iteration of the report was published in 2022, covering the years 2020 and 2021. Within the report, Meta details how it addresses human rights «concerns» associated with its products and services (G4b).
	Community Standards Enforcement Report	Each quarter, Meta publishes data on the actions it takes to restrict content violating its Community Standards for Instagram and Facebook, including policies on bullying and harassment, organized hate, and hate speech (F4a, F4b).
	Content Restrictions Based on Local Law	In this report, Meta provides data on the number of restrictions it complied with in response to government demands (F6). It covers Instagram and Facebook. The report includes a case studies page with examples on how the company responds to this type of demands (F5a).
	Intellectual Property (IP) report	This report details the number of IP violations Meta received and how much content they restricted, as a result, on Instagram and Facebook (F7).
TikTok	Intellectual Property Policy	TikTok outlines its Intellectual Property Policy and the legal basis under which it restricts content for violating intellectual property rights (F5b).
	How TikTok recommends videos #ForYou	In this policy, TikTok explains in detail how it uses algorithmic systems to recommend videos (F12).
Twitter	Terms of Service	A key policy governing users' access to and use of Twitter.
	Global Impact Report (2020)	The company's first ever Global Impact Report, which highlights the company's efforts to have a positive impact. It has not been published since 2020.
YouTube	AI principles	The document lays out a number of principles Google follows in developing and using AI, including that the systems are «socially beneficial» and do not lead to bias or discrimination.
	Ads Safety Report (2021)	Google's transparency report with data on actions taken against ads to «prevent malicious use of our ads platforms.»

Table 2

Our evaluation focused solely on the availability of policies and disclosures in the areas identified by the list of indicators listed above. We did not evaluate the language and terminology used in the Arabic policies to assess their clarity, comprehensiveness, or the presence of any discrepancies or translation errors compared to the English policies. However, other studies have attempted to examine substantively the content of policies in Arabic. For instance, a [2022 report](#) by Localization Lab and Internews titled “*Wait, Who’s Timothy McVeigh?*” evaluated the quality and usability of Facebook’s Community Standards and YouTube’s Community Guidelines in Arabic, as well as in three other languages (Amharic, Bengali, and Hindi). In their review of content moderation policies, they found that while Facebook’s policies were “readable,” the platform “routinely used literal translation and employed words that were unfamiliar and highly technical,” and “contextualized key policy concepts in terms most familiar to readers in Anglophone countries omitting explanations in terms familiar to readers in the MENA region.” Similarly, YouTube’s policies were considered “coherent,” but the company also used technical terms and “made reference to cultural internet phenomena most familiar to end-users in Anglophone countries, esp. the United States.”



Indicator-based Findings: Where Each Platform Stands

Human Rights Commitments, Due Diligence, and Content Moderation Appeals

All platforms, with the exception of TikTok, provide clear and explicit commitments to protect freedom of expression and information and the right to privacy (Indicator G1). TikTok failed to meet this criterion, having only published a general commitment to human rights that lacks a specific mention of freedom of expression, and its commitment to privacy was not grounded in international human rights standards as RDR's G1 indicator requires.

All platforms performed very poorly when it came to conducting human rights due diligence to identify and mitigate potential risks related to the enforcement of their policies related to freedom of expression and information, privacy, and the right to nondiscrimination (G6b). **TikTok, Twitter, and YouTube did not provide any disclosures regarding this type of due diligence in any of their markets, neither in Arabic nor in English.** Facebook, through its parent company Meta, offered incomplete disclosures about this type of due diligence in certain markets, such as the U.S. through an [independent civil rights audit](#) and in [Cambodia, Sri Lanka, Indonesia, Philippines, and Myanmar](#). However, Meta did not provide any evidence that it conducted such assessments in Lebanon.

Regarding platforms' processes for content moderation appeals (G6b), **only Facebook offered relatively clear and predictable processes and mechanisms for users to appeal content-moderation actions.** Its remedy policies included commitments to notify users affected by a content-moderation action, set timeframes for user notifications when such actions are taken, and clearly disclose the platform's process for reviewing appeals when they are submitted. However, Facebook failed to mention other important policies related to content moderation appeals, such as specific timeframes for reviewing appeals and the potential role of automation in these reviews. YouTube's policies in this area lacked clarity, as they did not always provide affected users with the ability to appeal all content-moderation actions or consistently notify them of such actions. TikTok and Twitter disclosed very little about their policies for content moderation appeals.

Enforcement of Policies

All U.S. platforms demonstrated relatively strong disclosures regarding their rules, including the content and activities they prohibit and how they enforce them (F3a). Among the platforms, **YouTube was the most transparent in this regard.** Twitter and YouTube were the only platforms that provided clear information on whether any government or private entities receive priority consideration when flagging content for potential restrictions based on rule violations. TikTok was the least transparent about this indicator, using vague language regarding reasons for restricting user accounts and providing general information about rule enforcement.

All platforms provided some level of data on content and account restrictions implemented to enforce their terms of service (F4a, b). However, none of them provided comprehensive data. TikTok, Twitter, and YouTube published data on content removal, but failed to account for other types of restrictions (F4a). For example, Twitter mentions in “[Our range of enforcement options](#)” that it employs methods beyond content removal, such as limiting visibility in users’ feeds, but it does not provide data on these types of restrictions in its [report](#). In their transparency reports, Twitter and TikTok included data on account removals, but did not cover other types of account restrictions (F4b). In its [Community Guidelines Enforcement report](#), YouTube reported the number of channels it removed, but did not specify the number of accounts affected. On the other hand, Facebook published aggregated data on actions taken to restrict both content and accounts in its [Community Standards Enforcement Report](#), but this data was not available in Arabic.

Algorithmic Systems and Bots

All platforms failed to disclose an explicit, clearly articulated policy commitment to human rights in their development and use of algorithmic systems (G1, E3). Meta and YouTube provided commitments that were not firmly grounded in human rights, while TikTok and Twitter did not make any commitments in this area. Google, in its “[AI principles](#)” policy, outlined several principles for AI development and utilization, including ensuring that the systems are “socially beneficial” and do not lead to bias or discrimination. However, it is not evident that human rights serve as the primary framework for Google’s governance of its AI development and algorithmic decision-making systems. On the other hand, Meta’s [Corporate Human Rights Policy](#) states: “Human rights also guide our work developing responsible innovation practices, including when building, testing, and deploying products and services enabled by Artificial Intelligence (AI).” The company also references the OECD Principles on Artificial Intelligence, acknowledging that: “We recognize the importance of the OECD Principles on Artificial Intelligence, which are widely adopted and endorsed by the G20.” However, this falls short of a clear commitment to human rights, as Meta only mentions being guided by and recognizing the importance of human rights.

Regarding the use of algorithmic systems to curate, recommend, and/or rank content (F12), TikTok provided the most details. It explicitly disclosed that it [uses](#) algorithmic systems for content curation, recommendation, and/or ranking, how these algorithmic systems are deployed, including the factors that influence these systems, user control options, and the default activation of these systems. **However, TikTok did not say anything about whether users can opt in to automated content curation, recommendation, and/or ranking systems.** Additionally, **TikTok failed to make any of this information available in Arabic**, which is problematic considering the limited understanding of Arabic-language algorithmic systems for content curation, recommendation, and/or ranking among researchers and journalists due to the lack of resources and tools. Making such information available in Arabic would enable researchers, journalists, and civil society groups in Lebanon and in the broader region to access the necessary information to hold platforms accountable for their deployment of algorithmic systems.

Facebook, Twitter, and YouTube provided similar levels of disclosures in both Arabic and English, although their policies were not as comprehensive as TikTok's English policies. **Facebook, only provided information about how its Feed curates and recommends content using algorithmic systems without specifying other features that use these algorithmic systems.** Feed is the list of content that is constantly being updated and appears on a [user's homepage](#) on Facebook. It [includes](#) "status updates, photos, videos, links, app activity and likes from people, Pages and groups that you follow on Facebook." Nevertheless, it is known that Facebook deploys algorithms in other ways as well, to [recommend](#) "friends" and show [search results](#) for instance.

Twitter clearly discloses how it uses algorithmic systems to curate, recommend, and/or rank the content that users can access through its platform. The disclosures include information about the variables that influence these systems. While it [provided](#) users with an option to "toggle between seeing the top Tweets first and the latest Tweets," Twitter failed to offer additional options for users to control these variables. The platform does not disclose whether algorithmic systems are used by default for automated content curation, recommendation, and/or ranking.

YouTube implies the use of algorithmic systems for content curation, recommendation, and/or ranking, but does not provide detailed information about how these systems work or the variables that influence them. While users have some options to control the variables that the algorithmic system takes into account, YouTube does not mention whether users can opt in to automated content curation, recommendation, and/or ranking systems.

Algorithmic transparency is essential for researchers, journalists, academics, civil society groups, policymakers, and other stakeholders to hold companies and platforms accountable for potential harms to user safety and human rights, including their right to nondiscrimination when developing and deploying algorithmic systems. Previous research and investigations have demonstrated how social media platforms' algorithmic systems, designed to drive engagement, can promote and spread hateful content by targeting it to users who are most likely to share it.

Finally, platforms lacked transparency regarding their policies governing the use of automated software agents, commonly known as "bots," and the enforcement of such policies.

Twitter was the most transparent in this regard, providing clear rules governing the use of bots on its platform in its "[Platform manipulation and spam policy](#)" and how the policy is enforced. However, the platform provided less information in Arabic about the enforcement of its bot policies.

TikTok and Meta provided minimal disclosures regarding their bot policies. TikTok disclosed policies about certain activities that can be generated by bots such as "Spam and fake engagement," in its [Community Guidelines](#). However, it did not disclose more comprehensive rules about bots in general and provide details on how it enforces such policies. Meta mentioned in its [Community Standards](#) that it prohibits "people to misrepresent themselves on Facebook, use fake accounts, artificially boost the popularity of content, or engage in behaviors designed to enable other violations under our Community Standards." However, this rule may not necessarily apply to all types of bots.

YouTube did not disclose any policies regarding the regulation of bots on its platform.

Targeted Advertising

Targeted advertising is the core of social media platforms' business models. Under the practice, third-party advertisers pay to show ads to users based on their personal data such as age, location, and interests.

All platforms disclosed ad content policies (F1b, F3b), with Twitter being the only platform that disclosed comprehensive policies, followed by YouTube. Twitter clearly outlined the types of advertising content it does not permit and the processes and technologies it uses to identify advertising content or accounts that violate its rules. It also required all advertising content be clearly labeled as such. YouTube also had clear ad content policies and disclosed how it enforces them, although it did not explicitly require all advertising content to be labeled. Facebook's ad content policies lacked clarity in both Arabic and English, failing to specify all prohibited types of ad content and the technologies used to enforce its ad content policies. TikTok was the least transparent, mentioning in its [General TikTok for Business Platform Terms](#) that it "may reject or remove Ad Materials or Ads at any time for any or no reason," and providing no indication whether or not it requires all ad content to be labeled as such.

Twitter and YouTube are equally leading in transparency about their ad targeting rules and how they enforce them (F3c). All platforms disclosed that they allow third-party advertisers to target users with advertising content. However, Twitter and YouTube provided more detailed information than Facebook and TikTok on prohibited targeting parameters and the processes and technologies used to identify advertising content or accounts that violate ad targeting rules. **All platforms disclosed that they allow advertisers to target specific individuals with ads using unique identifiers like email addresses, a problematic practice that often leads to privacy violations.**

Regarding transparency in actions taken to restrict advertising content, **TikTok and YouTube were the only platforms that published data on the volume and nature of actions taken** when the **advertising content violates the company's advertising content policies and advertising targeting policies in both English and Arabic (F4c).** They both reported the number of ads removed but did not disaggregate advertisements rejected for violating ad content rules from those rejected for violating ad targeting rules. Additionally, they did not provide a granular breakdown of the number of violations per specific rule.

Censorship Demands

U.S. platforms were transparent regarding their handling of government demands to restrict content and accounts, including responses to court orders and non-judicial government demands. They expressed a commitment to exercise due diligence on these demands before deciding to respond and pushing back on excessive requests. However, there were variations in the level of transparency among the platforms. Google and Twitter clearly disclosed the legal basis under which they may comply with

government demands, whereas Facebook's [content restrictions transparency report](#) provided some examples in its individual country reports of local laws under which it received requests to restrict content, but it did not provide a comprehensive list of the types of laws with which it may comply.

TikTok was the least transparent among all platforms in this area. Its policy for responding to these types of demands lacked details and mentioned no legal basis for complying with such demands.

All platforms were less transparent about their policies for handling private requests to restrict content and accounts compared to government demands (F5b). Their disclosures primarily focused on requests addressing copyright violations, overlooking other types of requests, such those pertaining to hate speech.

Twitter was the most transparent platform regarding censorship demands. It provided a clear legal basis for granting such demands, but the process for responding to all types of private requests was not clearly explained. While Twitter outlined its response process for private [copyright requests](#) and requests submitted under its [Child Sexual Exploitation Policy](#), it remained unclear how it handles [requests regarding hate speech](#) submitted by NGOs and other organizations under the [EU Code of conduct on countering illegal hate speech online](#).

YouTube disclosed a policy for responding to private requests related to [copyright infringement](#), but this policy lacked details. It remained unclear how the platform handles other types of private requests, and YouTube failed to disclose a clear commitment to carry out due diligence on this type of private requests and a commitment to push back against overboard requests. Similarly, Meta disclosed a policy for responding to private requests, but this policy does not cover all types of private requests, such as demands to remove hate speech content in partnership with the EU.

On this issue again, TikTok was outperformed by its U.S. peers. It disclosed minimal information about how it handles private requests for content and account restrictions. While TikTok [disclosed](#) a process for handling private Intellectual Property infringements requests and the legal basis under which it handles such requests, it remains unclear if the platform receives other types of private requests. It did not disclose a commitment to exercise due diligence on these requests before responding or pushing back against overboard requests.

In its [transparency report](#), YouTube provided the most data about government demands to restrict content and accounts (F6). The report listed the number of affected pieces of content or URLs, the subject matter associated with government demands, the legal authorities making the demands, and the number of demands complied with.

Twitter disclosed less information in its [transparency report](#) than YouTube but it was more transparent than TikTok and Facebook. While Twitter broke out the number of government demands received by country and listed the number of accounts affected, it did not provide details such as the number of affected pieces of content or URLs, or the types of subject matter associated with government demands it receives.

TikTok, on the other hand, outperformed Facebook in certain aspects, such as making its data available in Arabic, breaking out the number of government demands by country, and listing the number of accounts affected. Facebook did not provide this level of details. However, both platforms failed to disclose the types of subject matter associated with government demands they receive and the number of government demands from different legal authorities.

The review of data published by the four platforms since 2019 suggests that there have been relatively few censorship demands from Lebanese authorities, at least to YouTube, TikTok, and Twitter. However, the platforms do not provide sufficient information to understand the extent to which these demands are related to hate speech. This lack of information makes it difficult for civil society groups, digital rights advocates, journalists, and groups and initiatives advocating on behalf of minorities and those at most risk of hate speech in Lebanon to comprehend the actions taken by the Lebanese government and the platforms against hate speech and to protect freedom of expression. It also makes it challenging to hold the Lebanese government and the platforms accountable. This issue is particularly significant in a context like that of Lebanon, where authorities often restrict freedom of expression under broad reasons such as defamation.

Since 2019, Lebanon has [submitted](#) only one removal demand to Google affecting YouTube content. For TikTok, five demands were [submitted](#) between January and June 2021, resulting in the restriction of 48 accounts out of 55 targeted by the government due to violations of the platform's Community Guidelines. Twitter [received](#) four requests from Lebanese authorities in the second half of 2019, none of which it complied with. It is worth noting that Twitter has not published an updated transparency report since July 2022, a few months before Elon Musk took over the platform. Twitter's [latest report](#) covered the period from July to December 2021, and it is unclear if the company will resume publishing such reports under Musk's leadership, and if yes, whether the format and data provided will be different.

Regarding Facebook, the data provided is not comprehensive. It only includes the number of content restrictions without data on compliance rates, the number of demands received, and accounts restricted. Between January 2021 and June 2022, Facebook restricted 38 pieces of content in Lebanon that were "externally imposed" in [compliance](#) with "legal demands that assert extraterritorial jurisdiction." Additionally, between January 2019 and December 2020, 12 similar restrictions were reported, along with 43 pieces of content restricted in Lebanon specifically for defamation.

Overall, platforms provided less data on private requests to restrict content and accounts. TikTok and YouTube did not provide any data, while Meta and Twitter offered limited data in this area.

The table below summarizes the transparency level of platforms on the above-listed indicators:

	TikTok	Twitter	YouTube	Facebook
Human rights commitment	Lacks	Clear	Clear	Clear
Due diligence	Poor	Poor	Poor	Incomplete
Content moderation appeals	Poor	Limited	Limited	Relatively clear
Enforcement of policies	Vague	Clear	Clear	Clear
Algorithmic systems and bots	Limited	Limited	Limited	Limited
Targeted advertising	Lacks	Clear	Clear	Lacks
Censorship demands	Lacks	Limited	Limited	Limited

Table 3



How Platforms Define Hate Speech and Enforce Rules

A closer examination of these social media platforms’ policies enables a comparative analysis of their approaches to hate speech, discrimination, and content removal. **Among the four platforms, Twitter stands out as the only one that does not explicitly use the term “hate speech.”** Instead, it categorizes different types of hate speech under various headings, including “hateful conduct”, “violent speech,” and “violent and hateful entities.”

YouTube explicitly states that it does not allow hate speech, defining it as the promotion “of violence or hatred against individuals or groups based on any of the following attributes...” Facebook also provides a clear definition of hate speech as “a direct attack against people—rather than concepts or institutions—on the basis of what we call protected characteristics...” TikTok also offers a precise definition of hate speech as “content that attacks, threatens, incites violence against, or otherwise dehumanizes an individual or a group on the basis of the following protected attributes.”

All four companies establish a set of protected characteristics and attributes under these policies, including race, ethnicity, sexual orientation, religion/religious affiliation, among others. YouTube and ByteDance (TikTok) also include “immigration status” in their lists. Google (YouTube) and Twitter also have “age” in their respective lists. YouTube uniquely encompasses “victims of a major violent event and their kin” and “veteran status”. As a result, YouTube has the most comprehensive list of protected attributes and groups among the four platforms.

	Facebook	TikTok	Twitter	YouTube
Age			✓	✓
Caste	✓	✓	✓	✓
Disability	✓	✓	✓	✓
Ethnicity	✓	✓	✓	✓
Gender identity and/or Gender expression	✓	✓	✓	✓
Immigration status		✓		✓
National origin / Nationality	✓	✓	✓	✓
Race	✓	✓	✓	✓
Religion / Religious affiliation	✓	✓	✓	✓
Serious disease	✓	✓	✓	
Sex / Gender	✓	✓	✓	✓
Sexual orientation	✓	✓	✓	✓
Victims of a major violent event and their kin				✓
Veteran status				✓

Table 4

While YouTube and TikTok mention immigration status as a protected category in their hate speech policies, Facebook and Twitter do not. As previously demonstrated, refugees and migrant workers are among the most [vulnerable groups](#) in Lebanon, subjected to discrimination and hate speech both online and offline. Therefore, the exclusion of these groups as a protected category can further endanger them and expose them to violent online campaigns.

The other categories are comprehensive enough to encompass the main groups of people facing discrimination and hate speech in Lebanon. In addition to refugees and migrant workers, women, individuals with disabilities, and the LGBTQ+ community are also [vulnerable](#) to online and offline hate speech.

Nevertheless, platforms need to do more to address how their hate speech policies are enforced and make sure they are effectively implemented. This is particularly needed concerning their algorithms and the qualifications of the moderators they employ.

How Content Is Identified and Removed

All four platforms employ a combination of machine learning and human reviewers to filter through the content for violations of their policies and their ToS. Google (YouTube) and Meta (Facebook) are the most explicit platforms in terms of disclosing their removal and filtration processes that involve using machine learning. They are followed by Twitter, then ByteDance (TikTok). Additionally, all platforms provide a mechanism for users to appeal content removal or account closures.

YouTube utilizes explicit text and graphics to explain how information is input into their AI and machine learning systems, as well as [when exactly human](#) interventions come in, in the form of content reviewers. YouTube's machine learning systems automatically remove spam and previously removed content, while flagged content is first subjected to human review. Reviewers' input is continuously used to train and enhance YouTube's machine learning systems.

Meta (Facebook) also provides detailed explanations of how its AI systems and teams filter through content. It outlines that its systems can recognize images and understand text, and that its "integrity teams – who are responsible for scaling the detection and enforcement of our policies – build upon these models to create more specific models that make predictions about people and content. These predictions help us enforce our policies." Meta also adds that "for example, an AI model predicts whether a piece of content is hate speech or violent and graphic content. A separate system – our enforcement technology – determines whether to take an action, such as deleting, demoting or sending the content to a human review team for further review." Meta is [particularly clear](#) on the role of AI in detecting hate speech and how most of hate speech content is easily detectable even before individuals report it. The platform even extends its detection efforts to reactions and comments, searching for similarities with previously removed posts determined as hate speech.

Both TikTok and Twitter offer general information about the use of technologies and people to enforce their rules. Twitter mentions a “global team that manages enforcement of Twitter Rules 24/7 coverage of most supported languages on Twitter.” However, the details of this process are not readily available online.

Social media platforms are increasingly relying on algorithms in content moderation. While these technologies can accurately detect certain types of content like nudity, graphic imagery, or hateful terms in posts, their effectiveness decreases when applied to [situations](#) that require nuance and an understanding of the context to make the right judgment on whether a piece of content should remain online. Moreover, content moderation algorithms [lag behind](#) in languages other than English, raising questions about their robustness and accuracy in detecting hate speech in modern standard Arabic, Lebanese dialect, Anglo-Arabic, and Franco-Arabic (Arabic words transcribed into a combination of Latin script and Arabic numerals), which are widely used by social media users in Lebanon.

In such cases, adequately trained human content moderators who are familiar with the Lebanese context can help overcome the shortcomings of platform algorithms. They can minimize the removal of legitimate content and mitigate the risks of hateful content persisting online. Yet, time and again, media reports, research, and leaks have [shown](#) the insufficient investment by platforms in content moderation in the Global South, including Lebanon. For instance, documents [leaked](#) in 2021 by former Facebook data engineer turned whistleblower Frances Haughen revealed that Facebook was allocating 87 percent of its budget to counter misinformation to English-speaking users, despite representing only 9 percent of its user base.

Partnerships with Civil Society Organizations

Meta (Facebook), TikTok, and YouTube have content moderation partnerships with civil society organizations across the world. These partnerships, among other things, allow organizations to flag harmful content, including hate speech, to the platforms for removal. Users can also flag such content, but content flagged by partner organizations typically receive priority consideration.

Meta’s “[network of Trusted Partners](#) includes over 400 non-governmental organizations, humanitarian agencies, human rights defenders and researchers from 113 countries around the globe.” These partners not only flag problematic content on Facebook and Instagram but also “foster online safety and security,” and “inform the development of effective and transparent policies.” Meta provides examples of Trusted Partners, such as Tech4Peace in Iraq and Defy Hate Now in South Sudan, but does not publicly list the Lebanese organizations included in this network.

YouTube’s [Trusted Flagger program](#) equips government agencies and NGOs with tools to flag content that violates the platform’s community guidelines. However, YouTube does not disclose the names of its trusted flaggers, so it is unclear if it has any in Lebanon. The program also requires those participating to sign a non-disclosure agreement.

TikTok mentions that it works with [safety partners](#), including “industry experts, non-governmental organizations, and industry associations around the world in our commitment to building a safe platform for our community.” These partnerships cover areas such as fact-checking, “body inclusivity,” and countering violent extremism. Its safety partner in Lebanon is Embrace, the National Suicide Prevention Helpline, but it is unclear if TikTok has specific partners in the country addressing hate speech.

Finally, in December 2022, following Elon Musk’s takeover of Twitter, the platform [disbanded](#) its Trust and Safety Council. This council [brought together](#) a group of “independent expert organizations” to “advocate for safety and advise” Twitter in the development of their products, rules, and programs. One of the council members was SMEX, a Lebanese non-profit advocating for the protection of human rights in the digital space. SMEX had joined the council in 2017. Twitter notified member organizations via email that the council was no longer “the best structure” to bring “external insights into our product and policy development work.” This decision, along with other actions by Musk, such as mass layoffs affecting [content moderators](#) and reinstating previously banned accounts, has been criticized by civil society actors and former council members. These actions represent a setback to the progress the platform has made over the years in ensuring the safety of its users.

These developments are concerning for the safety of users and the openness of public debate, particularly in Lebanon. Social media platforms, including Twitter, are utilized by non-state actors and political groups in Lebanon to launch sectarian attacks and propagate hateful campaigns against their opponents and critics. For instance, during the 2022 Lebanese election period, hateful [campaigns](#) targeted Dalia Ahmad, a black television host of Sudanese origin working with Al Jadeed TV. Ahmad had criticized Lebanese politicians, including Hezbollah leader Hassan Nasrallah and then President Michel Aoun, describing them as “crocodiles” during her TV show. Subsequently, racist and hateful comments and messages targeting her spread on Facebook and Twitter.



Recommendations for Platforms

To address the challenges related to hate speech and content moderation in Lebanon, social media platforms should take the following actions:

- **Invest adequate resources in content moderation in Lebanon and the wider MENA and Arabic-speaking region.** Platforms should prioritize hiring qualified moderators who understand the Lebanese and wider Middle Eastern contexts and local languages. This will help mitigate the shortcomings of algorithms and reduce the risks of removing legitimate content while allowing hateful content to persist.
- **Adequately test algorithmic systems and assess risks associated with their use.** Platforms should ensure that algorithmic systems, used to moderate, rank, curate, or recommend content, are rigorously tested on diverse datasets that include Arabic and its various dialects and variations.
- **Provide appropriate and effective tools for appealing content moderation decisions.** Users should be able to appeal when their content is taken down or when their reports of harmful content are rejected. In cases where algorithms are responsible for content moderation actions, users should be able to appeal to a human moderator for a fair review.
- **Conduct human rights impact assessments in Lebanon and other divided societies in the region.** Assessments should address risks posed by platforms' enforcement of their processes and their use of algorithmic systems (for instance, to moderate and to rank/recommend content) to users' fundamental rights, including freedom of expression and non-discrimination. These assessments should particularly look into how these policies affect vulnerable communities and at-risk groups, who often face the brunt of discrimination and hate speech, such as Palestinian and Syrian refugees, women, migrant workers, and queer individuals and communities.
- **Improve data transparency on government demands to restrict content and accounts.** Meta, in particular, should break down the number of government demands it receives by country, specify the number of accounts affected, and publish clear compliance rates. Along with TikTok, Meta should also detail the subject matter associated with these demands, including those related to hate speech. All platforms should include examples of demands they receive from Lebanon, especially those pertaining to hate speech.
- **Publish data on private requests to restrict hate speech content.** Such data should include requests submitted under the "2016 EU Code of conduct on countering illegal hate speech online" and through other processes or partnerships, including those submitted by private entities in Lebanon, such as NGOs.
- **Strengthen partnerships with Lebanese NGOs to address hate speech.** Such partnerships should go beyond mere public relations stunts and actively address the concerns raised by trusted human rights groups and civil society organizations.

By working together, platforms and NGOs can develop effective solutions to tackle hate speech on social media. While expanding trusted flagging mechanisms can be beneficial, it is crucial to ensure they do not serve as a smoke screen to hide the core issue: Without radical changes to the design, policies, and practices of platforms, hate speech will remain a threat to users and communities in Lebanon and around the world.

Appendix 1: List of Key Terms and Definitions

Account restriction / restrict a user's account: Limitation, suspension, deactivation, deletion, or removal of a specific user account or permissions on a user's account.

Advertising audience categories: Groups of users, identified for the purpose of delivering targeted advertising, who share certain characteristics and/or interests, as determined on the basis of user information that a company has either collected or inferred.

Advertising content: Any content that someone has paid a company to display to its users.

Advertising content policies: Documents that outline a company's rules governing what advertising content is permitted on the platform.

Advertising targeting policies: Documents that outline a company's rules governing what advertising targeting parameters are permitted on the platform.

Algorithms: An algorithm is a set of instructions used to process information and deliver an output based on the instructions' stipulations. Algorithms can be simple pieces of code, but they can also be incredibly complex, "encoding for thousands of variables across millions of data points." In the context of internet, mobile, and telecommunications companies, some algorithms — because of their complexity, the amounts and types of user information fed into them, and the decision-making function they serve — have significant implications for users' human rights, including freedom of expression and privacy. See more at: ["Algorithmic Accountability: A Primer," Data & Society](#).

Algorithmic system: A system that uses algorithms, machine learning and/or related technologies to automate, optimize and/or personalize decision-making processes.

Algorithmic system development policies: Documents that outline a company's practices related to the development and testing of algorithms, machine learning and automated decision-making.

Appeal: For RDR's purposes, this definition of appeals includes processes through which users request a formal change to a content moderation or account restriction decision made by a company.

Bot: An automated online account where all or substantially all of the actions or posts of that account are not the result of a person.

Clearly disclose(s): The company presents or explains its policies or practices in its public-facing materials in a way that is easy for users to find and understand.

Content: The information contained in wire, oral, or electronic communications (e.g. a conversation that takes place over the phone or face-to-face, the text written and transmitted in an SMS or email).

Content moderation action: Content moderation actions are steps platforms take to restrict the visibility of content or the capabilities of a user account. They may be performed by humans, automated systems, or a mix of both.

Content restriction: An action the company takes that renders an instance of user-generated content invisible or less visible on the platform or service. This action could involve removing the content entirely or take a less absolute form, such as hiding it from only certain users (e.g. inhabitants of some country or people under a certain age), limiting users' ability to interact with it (e.g. making it impossible to "like"), adding counter speech to it (e.g. corrective information on anti-vaccine posts), or reducing the amount of amplification provided by the platform's curation systems.

Court orders: Orders issued by a court. They include court orders in criminal and civil cases.

Curate, recommend, and/or rank: The practice of using algorithms, machine learning and other automated decision-making systems to manage, shape, and govern the flow of content and information on a platform, typically in a way that is personalized to each individual user.

Flag: The process of alerting a company that a piece of content or account may be in violation of the company's rules, or the signal that conveys this information to the company. This process can occur either within the platform or through an external process. Flaggers include users, algorithmic systems, company staff, governments, and other private entities.

Human Rights Impact Assessments (HRIA): HRIsAs are a systematic approach to due diligence. A company carries out these assessments or reviews to see how its products, services, and business practices affect the freedom of expression and privacy of its users.

Non-judicial government demands: These are requests that come from government entities that are not judicial bodies, judges, or courts. They can include requests from government ministries, agencies, police departments, police officers (acting in official capacity), and other non-judicial government offices, authorities, or entities.

Policy commitment: A publicly available statement that represents official company policy which has been approved at the highest level of the company.

Private requests: Requests made through a private process rather than a judicial or governmental process. Private requests for content restriction can come from a self-regulatory body such as the Internet Watch Foundation, or a notice-and-takedown system, such as the U.S. Digital Millennium Copyright Act.

Targeting parameters: The conditions, typically set by the advertiser, that determine which users will be shown the advertising content in question. This can include users' demographics, location, behavior, interests, connections, and other user information.

Terms of service: This document may also be called Terms of Use, Terms and Conditions, etc. The terms of service “often provide the necessary ground rules for how various online services should be used,” as stated by the EFF, and represent a legal agreement between the company and the user. Companies can take action against users and their content based on information in the terms of service. Source: [Electronic Frontier Foundation, “Terms of \(Ab\)use”](#).

Third party: A party or entity that is anything other than the user or the company. For the purposes of this methodology, third parties can include government organizations, courts, or other private parties (e.g. a company, an NGO, an individual person).

Appendix 2: Research Indicators and Elements

G1. Policy Commitment

The company should publish a formal **policy commitment** to respect users’ human rights to freedom of expression and information and privacy.

Elements:

1. Does the company make an **explicit**, clearly articulated **policy commitment** to human rights, including to freedom of expression and information?
2. Does the company make an **explicit**, clearly articulated **policy commitment** to human rights, including to privacy?
3. Does the company disclose an **explicit**, clearly articulated **policy commitment** to human rights in its development and use of **algorithmic systems**?

G4(b). Impact assessment: Processes for policy enforcement

The company should conduct regular, comprehensive, and credible due diligence, such as through robust **human rights impact assessments**, to identify how its processes for policy enforcement affect users’ fundamental rights to freedom of expression and information, to privacy, and to non-discrimination, and to mitigate any risks posed by those impacts.

Elements:

1. Does the company **assess** freedom of expression and information risks of enforcing its terms of service?
2. Does the company conduct **risk assessments** of its enforcement of its privacy policies?
3. Does the company **assess** discrimination risks associated with its processes for enforcing its **terms of service**?

4. Does the company **assess discrimination** risks associated with its processes for enforcing its **privacy policies**?
5. Does the company conduct additional evaluation whenever the company's **risk assessments** identify concerns?
6. Do **senior executives** and/or members of the company's **board of directors** review and consider the results of **assessments** and due diligence in their decision-making?
7. Does the company conduct **assessments** on a regular schedule?
8. Are the company's **assessments** assured by an external **third party**?
9. Is the external **third party** that assures the **assessment** accredited to a relevant and reputable human rights standard by a credible organization?

G6(b). Process for content moderation appeals

The company should offer users clear and predictable **appeals** mechanisms and processes for appealing **content-moderation actions**.

Elements:

1. Does the company **clearly disclose** that it offers **affected users** the ability to **appeal content-moderation actions**?
2. Does the company **clearly disclose** that it **notifies** the users who are **affected** by a **content-moderation action**?
3. Does the company **clearly disclose** a timeframe for **notifying affected users** when it takes a **content-moderation action**?
4. Does the company **clearly disclose** when **appeals** are not permitted?
5. Does the company **clearly disclose** its process for reviewing **appeals**?
6. Does the company **clearly disclose** its timeframe for reviewing **appeals**?
7. Does the company **clearly disclose** that such appeals are reviewed by at least one human not involved in the original **content-moderation action**?
8. Does the company **clearly disclose** what role automation plays in reviewing **appeals**?
9. Does the company **clearly disclose** that the **affected users** have an opportunity to present additional information that will be considered in the review?
10. Does the company **clearly disclose** that it provides the **affected users** with a statement outlining the reason for its decision?
11. Does the company **clearly disclose** evidence that it is addressing content moderation **appeals**?

F1(a). Access to terms of service

The company should offer **terms of service** that are **easy to find** and **easy to understand**.

Elements:

1. Are the company's **terms of service** **easy to find**?
2. Are the **terms of service** available in Arabic, the primary language spoken by users in Lebanon?
3. Are the **terms of service** presented in an **understandable manner**?

F1(b). Access to advertising content policies

The company should offer **advertising content policies** that are **easy to find** and **easy to understand**.

Elements:

1. Are the company's **advertising content policies** **easy to find**?
2. Are the company's **advertising content policies** available in the primary language(s) spoken by users in the company's home jurisdiction?
3. Are the company's **advertising content policies** presented in an **understandable manner**?

F1(c). Access to advertising targeting policies

The company should offer **advertising targeting policies** that are **easy to find** and **easy to understand**.

Elements:

1. Are the company's **advertising targeting policies** **easy to find**?
2. Are the **advertising targeting policies** available in Arabic, the primary language spoken by **users** in Lebanon?
3. Are the **advertising targeting policies** presented in an **understandable manner**?

F3(a). Process for terms of service enforcement

The company should **clearly disclose** the circumstances under which it may restrict **content** or **user accounts**.

Elements:

1. Does the company **clearly disclose** what types of **content** or activities it does not permit?
2. Does the company **clearly disclose** why it may **restrict a user's account**?

3. Does the company **clearly disclose** information about the processes it uses to identify **content** or **accounts** that violate the company's rules?
4. Does the company **clearly disclose** how it uses **algorithmic systems** to flag **content** that might violate the company's rules?
5. Does the company **clearly disclose** whether any government authorities receive priority consideration when **flagging content** to be restricted for violating the company's rules?
6. Does the company **clearly disclose** whether any private entities receive priority consideration when **flagging content** to be restricted for violating the company's rules?
7. Does the company **clearly disclose** its process for enforcing its rules once violations are detected?

F3(b). Advertising content rules and enforcement

The company should **clearly disclose** its policies governing what types of advertising content is prohibited.

Elements:

1. Does the company **clearly disclose** what types of **advertising content** it does not permit?
2. Does the company **clearly disclose** whether it **requires** all **advertising content** be clearly labelled as such?
3. Does the company **clearly disclose** information about the processes and technologies it uses to identify **advertising content** or **accounts** that violate the company's rules?

F3(c). Advertising targeting rules and enforcement

The company should **clearly disclose** its policies governing what type of **advertising targeting** is prohibited.

Elements:

1. Does the company **clearly disclose** whether it enables **third parties** to target its **users** with **advertising content**?
2. Does the company **clearly disclose** what types of **targeting parameters** are not permitted?
3. Does the company **clearly disclose** that it does not permit **advertisers** to target specific individuals?
4. Does the company **clearly disclose** that **algorithmically** generated **advertising audience categories** are evaluated by human reviewers before they can be used?

5. Does the company **clearly disclose** information about the processes and technologies it uses to identify **advertising content** or **accounts** that violate the company's rules?

F4(a). Data about content restrictions to enforce terms of service

The company should **clearly disclose** and regularly publish data about the volume and nature of actions taken to **restrict content** that violates the company's rules.

Elements:

1. Does the company publish data about the total number of pieces of **content restricted** for violating the company's rules?
2. Does the company publish data on the number of pieces of **content restricted** based on which rule was violated?
3. Does the company publish data on the number of pieces of **content** it restricted based on the format of content? (e.g. text, image, video, live video)?
4. Does the company publish data on the number of pieces of **content** it **restricted** based on the method used to identify the violation?
5. Does the company publish this data at least four times a year?
6. Can the data be exported as a **structured data** file?

F4(b). Data about account restrictions to enforce terms of service

The company should **clearly disclose** and regularly publish data about the volume and nature of actions taken to **restrict accounts** that violate the company's rules.

Elements

1. Does the company publish data on the total number of **accounts restricted** for violating the company's own rules?
2. Does the company publish data on the number of **accounts restricted** based on which rule was violated?
3. Does the company publish data on the number of **accounts restricted** based on the method used to identify the violation?
4. Does the company publish this data at least four times a year?
5. Can the data be exported as a **structured data** file?

F4(c). Data about advertising content and advertising targeting policy enforcement

The company should **clearly disclose** and regularly publish data about the volume and nature of actions taken to **restrict advertising content** that violates the company's **advertising content policies** and **advertising targeting policies**.

Elements:

1. Does the company publish the total number of **advertisements** it **restricted** to enforce its **advertising content policies**?
2. Does the company publish the number of **advertisements** it **restricted** based on which **advertising content** rule was violated?
3. Does the company publish the total number of **advertisements** it **restricted** to enforce its **advertising targeting policies**?
4. Does the company publish the number of **advertisements** it **restricted** based on which **advertising targeting rule** was violated?
5. Does the company publish this data at least once a year?
6. Can the data be exported as a **structured data file**?

F5(a). Process for responding to government demands to restrict content or accounts

The company should **clearly disclose** its process for responding to **government demands**—(including judicial orders) to remove, filter, or restrict **content** or **accounts**.

Elements:

1. Does the company **clearly disclose** its process for responding to **non-judicial government demands**?
2. Does the company **clearly disclose** its process for responding to **court orders**?
3. Does the company **clearly disclose** its process for responding to **government demands** from foreign jurisdictions?
4. Do the company's explanations **clearly disclose** the legal basis under which it may comply with **government demands**?
5. Does the company **clearly disclose** that it carries out due diligence on **government demands** before deciding how to respond?
6. Does the company commit to push back on inappropriate or overbroad **demands made by governments**?
7. Does the company provide clear guidance or examples of implementation of its process of responding to **government demands**?

F5(b). Process for responding to private requests for content or account restriction

The company should **clearly disclose** its process for responding to **requests** to remove, filter, or restrict **content** or **accounts** that come through **private processes**.

Elements:

1. Does the company **clearly disclose** its process for responding to **requests** to remove, filter, or restrict **content** or **accounts** made through **private processes**?
2. Do the company's explanations **clearly disclose** the basis under which it may comply

with **requests** made through **private processes**?

3. Does the company **clearly disclose** that it carries out due diligence on **requests** made through **private processes** before deciding how to respond?
4. Does the company commit to push back on inappropriate or overbroad **requests** made through **private processes**?
5. Does the company provide clear guidance or examples of implementation of its process of responding to **requests** made through **private processes**?

F6. Data about government demands to restrict for content and accounts

The company should regularly publish data about **government demands** (including judicial orders) to remove, filter, or restrict **content** and **accounts**.

Elements:

1. Does the company break out the number of **demands** it receives by country?
2. Does the company list the number of **accounts** affected?
3. Does the company list the number of pieces of **content** or URLs affected?
4. Does the company list the types of subject matter associated with the **demands** it receives?
5. Does the company list the number of **demands** that come from different legal authorities?
6. Does the company list the number of **demands** it knowingly receives from government officials to restrict **content** or **accounts** through **unofficial processes**?
7. Does the company list the number of **demands** with which it complied?
8. Does the company publish the original **demands** or disclose that it provides copies to a **public third-party archive**?
9. Does the company report this data at least once a year?
10. Can the data be exported as a **structured data** file?

F7. Data about private requests for content or account restriction

The company should regularly publish data about requests to remove, filter, or restrict access to **content** or **accounts** that come through **private processes**.

Elements:

1. Does the company break down the number of requests to restrict **content** or **accounts** that it receives through **private processes**?

2. Does the company list the number of **accounts** affected?
3. Does the company list the number of pieces of **content** or URLs affected?
4. Does the company list the reasons for removal associated with the requests it receives?
5. Does the company **clearly disclose** the **private processes** that made requests?
6. Does the company list the number of requests it complied with?
7. Does the company publish the original requests or disclose that it provides copies to a **public third-party archive**?
8. Does the company report this data at least once a year?
9. Can the data be exported as a **structured data** file?
10. Does the company **clearly disclose** that its reporting covers all types of requests that it receives through **private processes**?

F12. Algorithmic content curation, recommendation, and/or ranking systems

Companies should **clearly disclose** how users' online **content** is **curated, ranked, or recommended**.

Elements:

1. Does the company **clearly disclose** whether it uses **algorithmic systems** to **curate, recommend, and/or rank** the **content** that **users** can access through its platform?
2. Does the company **clearly disclose** how the **algorithmic systems** are deployed to **curate, recommend, and/or rank content**, including the variables that influence these systems?
3. Does the company **clearly disclose** what options users have to control the variables that the **algorithmic content curation, recommendation, and/or ranking system** takes into **account**?
4. Does the company **clearly disclose** whether **algorithmic systems** are used to automatically **curate, recommend, and/or rank content** by default?
5. Does the company **clearly disclose** that users can opt in to automated **content curation, recommendation, and/or ranking systems**?

F13. Automated software agents ("bots")

Companies should **clearly disclose** policies governing the use of **automated software agents ("bots")** on their platforms, products and services, and how they enforce such policies.

Elements:

1. Does the company **clearly disclose** rules governing the use of **bots** on its platform?

2. Does the company **clearly disclose** that it requires **users** to clearly label all **content** and **accounts** that are produced, disseminated or operated with the assistance of a **bot**?
3. Does the company **clearly disclose** its process for enforcing its **bot policy**?
4. Does the company **clearly disclose** data on the volume and nature of user **content** and **accounts restricted** for violating the company's **bot policy**?

P1(b). Access to algorithmic system development policies

The company should offer **algorithmic system development policies** that are **easy to find** and **easy to understand**.

Elements:

1. Are the company's **algorithmic system development policies easy to find**?
2. Are the **algorithmic system development policies** available in the primary language(s) spoken by users in the company's home jurisdiction?
3. Are the **algorithmic system development policies** presented in an **understandable manner**?

You are free to share, copy, distribute, and transmit this work under the conditions that you attribute the work to the Samir Kassir Foundation and Ranking Digital Rights but without suggesting in any way that the Samir Kassir Foundation and Ranking Digital Rights endorse you or your use of the work. You may not use this work for commercial purposes.

July 2023

Samir Kassir Foundation
Address: Riverside Bloc C, 6th floor, Charles Helou Street,
Sin el-Fil, Metn - Lebanon
Tel: +961 1 499012
Email: info@skeyesmedia.org
www.skeyesmedia.org

Ranking Digital Rights
Address: 740, 15th Street, N.W., Suite 900
Washington, DC 20005
Email: info@rankingdigitalrights.org
www.rankingdigitalrights.org