Twitter
Turkey Transparency Report
December 2022

Overview

Twitter was founded on a commitment to transparency. This commitment is part of our effort to serve the public conversation and to increase its collective health, openness, and civility around the world.

Twitter is reflective of real conversations happening in the world and that sometimes includes perspectives that may be offensive, controversial, and/or bigoted to others. While we welcome everyone to express themselves in our service, we will not tolerate behavior that harasses, threatens, dehumanizes or uses fear to silence the voices of others. We have the Twitter Rules in place to help ensure everyone feels safe expressing their beliefs and we strive to enforce them with uniform consistency. We are continually working to update, refine, and improve both our enforcement and our policies, informed by in-depth research around trends in online behavior both on and off Twitter, feedback from the people who use Twitter, and input from a number of external entities, including members of our Trust & Safety Council.[1] When it comes to enforcing these rules, we are committed to being fair, informative, responsive, and accountable. Read more about our approach to policy development and enforcement philosophy in the Twitter Help Center.[2]

We have a global team that manages enforcement of the Twitter Terms of Service and our Rules with 24/7 coverage in every supported language on Twitter. Our goal is to apply the Twitter Rules objectively and consistently. Enforcement actions are taken on content that is determined to violate the Twitter Rules.

The Twitter Rules, cover violence, terrorism/violent extremism, child sexual exploitation, abuse/harassment, hateful conduct, promoting suicide or self-harm, sensitive media (including graphic violence and adult content), and illegal or certain regulated goods or services. More information about each policy can be found in the Twitter Rules.

[1] https://blog.twitter.com/en_us/topics/company/2019/rules-refresh.html
[2] https://help.twitter.com/de/rules-and-policies/enforcement-philosophy

Twitter
Turkey Transparency Report
December 2022

Twitter Representative in Turkey & How to Report Violations

Twitter users in Turkey can reach out to the Twitter representative to report possible violation of their personal rights and privacy under Law No. 5651 through the following contact information:

Name of the Entity: Twitter İnternet Içerik Hizmetleri Limited Şirketi

Address: Esentepe Mah Büyükdere Cad Kanyon Blok No: 185 İç Kapı No: 271 Şişli / İstanbul

Twitter users in Turkey can also file reports of possible violations of the Twitter Rules in a variety of ways that are described in detail on the Twitter Help Center page.

Information on specific features on Twitter

Hashtags[3]

A hashtag—written with a # symbol—is used to index keywords or topics on Twitter. This function was created on Twitter, and allows people to easily follow topics they are interested in. If you Tweet with a hashtag on a public account, anyone who does a search for that hashtag may find your Tweet. Users can type a hashtagged keyword in the search bar to discover content and accounts based on their interests.

Algorithms that reduce or increase visibility[4]

When we take enforcement actions, we may do so either on a specific piece of content (e.g., an individual Tweet or Direct Message) or on an account. A few of the ways in which we might take action include limiting Tweet visibility. This makes content less visible on Twitter, either by making Tweets ineligible for amplification in Top search results and on timelines for users who don't follow the Tweet author, by downranking Tweets in replies (except when the user follows

[3] How to use hashtags: https://help.twitter.com/en/using-twitter/how-to-use-hashtags
[4] Our range of enforcement options: https://help.twitter.com/en/rules-and-policies/enforcement-options

the Tweet author), and/or excluding Tweets and/or accounts in email or in-product recommendations. Limiting Tweet visibility depends on a number of signals about the nature of the interaction and the type of the content.

Methods for the automated detection of content to be removed

There are many measures that have been in place for a long time on Twitter that relate, amongst other things, to the mitigation of child sexual exploitation and terrorism activity. Twitter does not tolerate any material that features or promotes child sexual exploitation — whether in Direct Messages or elsewhere throughout the service. This includes media, text, illustrations, computer-generated images, or the use of our service to advertise such material on other platforms. Twitter actively participates in the Tech Coalition to collaborate on the best practices to prevent and disrupt the spread of child sexual exploitation material. When we remove content, we immediately report it to the National Center for Missing and Exploited Children (NCMEC). NCMEC makes reports available to the appropriate law enforcement agencies around the world to facilitate investigations and prosecutions. Twitter was one of the founding members of, and continues to participate in, the Global Internet Forum to Counter Terrorism. A vast majority of all accounts that are suspended for the promotion of terrorism and child sexual exploitation are proactively flagged by a combination of technology and other purpose-built internal proprietary tools. You can learn more about our commitment to eradicating child sexual exploitation and terrorist content and the actions we've taken here. Our continued investment in proprietary technology is steadily reducing the burden on people to report to us.

Twitter employs a combination of heuristics and machine learning algorithms to automatically detect content that violates the Twitter rules and policies enforced on our platform.[5]

Heuristics are typically utilized to react quickly to new forms of violations that emerge on the platform. Heuristics are commonly patterns of text or keywords that may be typical of a certain category of violations. Most pieces of content detected by heuristics get reviewed by human agents before an action has been taken on the content.

Machine learning models vary in complexity and in the outputs they produce. For example, the model used to detect abuse on the platform is trained on abuse violations detected in the past.

[5] Twitter does not use automated detection for withholding procedures.

A majority of the content flagged by these machine learning models are also reviewed by human agents before an action is taken.

The nature of machine learning models used varies significantly by application. For automated detection of violations of policies in tweets, we use combinations of natural language processing models, image processing models and other sophisticated machine learning methods. For example, to detect images containing sensitive content such as gore or nudity, we use machine learning models that work with images as inputs.

Heuristics and Machine learning models are trained on thousands of datapoints with labels (e.g. violative or not) generated by trained human agents. Inputs to the models include the text within the tweet itself, the images attached to the tweet, and other features. Training data comes from both the cases reviewed by our content moderators, and a random sample of pieces of content from the platform.

In addition, our current methods of surfacing potentially violating content for review include leveraging the shared industry hash database, e.g. supported by the Global Internet Forum to Counter Terrorism (GIFCT), and deploying a range of internal tools and/or utilizing the industry hash sharing (e.g., PhotoDNA) prior to any reports filed. We commit to continuing to invest in technology that improves our capability to detect and remove e.g. terrorist and violent extremist content online, including the extension or development of digital fingerprinting and AI based technology solutions. Our participation in multi stakeholder communities, such as Global Internet Forum to Counter Terrorism (GIFCT) and the Christchurch Call to Action, helps to identify emerging trends in how terrorists and violent extremists are using the Internet to promote their content and exploit online platforms.

Automated policy enforcement (moderation algorithms) under the Twitter rules and policies undergoes rigorous testing before being applied to the consumer product and operating at Twitter scale; and once deployed, regular checks occur to ensure expected performance standards are being met. Before any given algorithm is launched to the platform, we verify its detection of policy violating content or behavior by drawing a statistically significant test sample and performing item-by-item human review. Reviewers have expertise in the applicable policies and training by our Policy teams to ensure the reliability of their decisions. During this testing phase we also calculate the expected volume of moderation actions a given automation is likely to perform in order to set a baseline against which we can monitor for anomalies in the future

(we call this “sizing”). Human review helps us to confirm that automations achieve a level of precision, and sizing helps us understand what to expect once the automations are launched. Once reviewers have confirmed that the detection meets an acceptable standard of accuracy, we consider the automation to be ready for launch. Once launched, automations are monitored dynamically for performance and health. If we detect anomalies in performance (for instance, significant spikes or dips against the volume we established during sizing), Engineering, Data Science and Policy teams revisit the automation to diagnose any potential problems, changing the automations as appropriate.

We’re always striving to work in a way that’s transparent and easy to understand, but we don’t always get this right. In October 2020, we heard feedback from people on Twitter that our image cropping algorithm didn’t serve all people equitably. As part of our commitment to address this issue, we also shared that we'd analyze our model again for bias. Over the last years, our teams have accelerated improvements for how we assess algorithms for potential bias and improve our understanding of whether machine learning is always the best solution to the problem at hand. In May 2021, we shared the outcomes of our bias assessment and a link for those interested in reading and reproducing our analysis in more technical detail.

In the context of our Responsible Machine Learning Initiative, we’re committed to sharing our learnings and asking for feedback. Both inside and outside of Twitter, we share our learnings and best practices to improve the industry’s collective understanding of this topic, help us improve our approach, and hold us accountable. This may come in the form of peer-reviewed research, data-insights, high-level descriptions of our findings or approaches, and even some of our unsuccessful attempts to address these emerging challenges. We’ll continue to work closely with third party academic researchers to identify ways we can improve our work and encourage their feedback. For further details on the access of information to science and research please see section “Access to information to scientific and research circles” below.

Advertisement and Paid Partnership Policies[6]

Tweets promoted through Twitter's advertising services are labeled as “Promoted” and must abide by our Twitter Ads Policies. Organic, non-promoted Tweets may also be considered paid product placements, endorsements, or advertisements (‘Paid Partnerships”). Advertisements posted as organic Tweets will require disclosures to viewers indicating the commercial nature of

[6] About rules and best practices with account behaviors:
https://help.twitter.com/en/rules-and-policies/twitter-rules-and-best-practices

such content. In addition to abiding by the Twitter Rules, users, including creators and brands, that participate in Paid Partnerships are responsible for complying with all applicable laws and regulations.

Transparency Policies[7]

Defending and respecting the user's voice is one of our core values at Twitter. Transparency is also an important part of this commitment. Some examples of steps we take to defend and respect our users include:

- Our transparency report — published bi-annually since 2012;
- Our user notice policies;
- Challenges to court orders to remove content from our platform or disclose user data;
- Publishing content removal demands to Lumen (aka Chilling Effects);
- Providing notices when content is removed or withheld or accounts suspended;
- Giving people access to their account information;
- Publishing a range of resources for people and organizations using our service, the public and law enforcement
- Sharing clear guidelines about appropriate uses of Twitter's Public APIs and Gnip data products

While this is a whole company effort, we also have a Trust and Safety team whose mandate is the protection of the people and organizations using our service and building trust in Twitter.

Updating your privacy and content preferences

Users can change your custom settings to control more of your Twitter experience. This includes privacy and safety preferences, the notifications you receive, display settings, and more. You can find more on this here[8]. Learn how to stay safe by managing your privacy settings on this

[7] Defending and respecting the rights of people using our service: https://help.twitter.com/en/rules-and-policies/defending-and-respecting-our-users-voice
[8] https://help.twitter.com/en/resources/twitter-guide/topics/how-to-make-the-best-twitter-for-you/how-to-use-custom-settings-on-twitter-twitter-help

page[9] in our Help Center. You are in control of how much information you share on Twitter and you can visit this page[10] to read more about it.

Mechanisms for Reviewing Reports and Team Information

Each complaint we receive is reviewed under Twitter's TOS and Rules. Any content we determine to violate Twitter's TOS and/or Rules is completely removed from the service. Twitter has built a team made up of individuals who have received training in order to handle these reports. This team consists of different tier groups, with higher tiers consisting of more senior, or more specialised, individuals. All members of the team involved in solving Turkish reports are fluent in Turkish and are required to undertake a language proficiency test in Turkish.

These individuals take appropriate action on the content after carefully reviewing the report and available context in close detail. If the content is not manifestly violative, it can be escalated for second or third opinions to policy specialists or internal legal teams. Everyone involved in this process works closely together with regular exchanges through meetings and other channels to ensure the timely and accurate handling of these reports.

Furthermore, all teams involved in handling these reports closely collaborate with a variety of other policy teams at Twitter who focus on safety policies, site integrity, or policies related to cybercrimes. This cross-team effort is particularly important in the aftermath of tragic events, such as violent attacks, to ensure alignment and swift action on violative content happens.

The team is supported by leads, subject matter experts, quality auditors and trainers. The team that handles these reports coming from users in Turkey have educational backgrounds that vary, but with the majority of the team holding an advanced qualification e.g Bachelors or Masters degrees. We ensure we hire people with diverse backgrounds in fields such as law, political science, psychology, communications, business, and languages. Team members who handle these reports in Turkish are all fluent in Turkish and English, with some agents speaking additional languages, including Bulgarian, German, Kurdish and Russian.

[9] https://help.twitter.com/en/safety-and-security#ads-and-data-privacy
[10] https://help.twitter.com/en/safety-and-security/twitter-privacy-settings

Twitter
Turkey Transparency Report
December 2022

All team members working on these reports are trained and retrained regularly on our policies, including sessions on cultural and historical context. Initially when joining the team at Twitter, each individual follows an onboarding program and receives individual mentoring during this period. Employees are rigorously trained on Twitter's TOS, Twitter's Rules, and local context as well as the internal tools and processes required for handling such complaints.

Employees have direct access to robust training and workflow documentation for the entirety of their employment, and are able to seek guidance at any time from trainers, leads, and internal specialist legal and policy teams as outlined above.

Updates about significant current events or policy changes are shared with all agents in real time, to give guidance and facilitate balanced and informed decision making. Calibration sessions are frequently carried out, focussing on different policies and offering clarifications regarding market trends or other questions raised by the reviewers. These sessions aim to increase collective understanding and focus on the needs of the agents in their day-to-day work.

The entire team also participates in obligatory TOS and Twitter Rules refresher trainings, as the need arises, or whenever policies are updated. These trainings are delivered by the relevant policy specialists who were directly involved in the development of the policy change. For these sessions we also employ the "train the trainer" method to ensure timely training delivery to the whole team across all of the shifts. All team members use the same training materials to ensure consistency.

In addition, given the nature and sensitivity of their work, the entire team has access to online resources and regular onsite group- and 1:1-sessions related to resilience and well-being. These are provided by mental health professionals. The team also attends resilience, self-care, and vicarious trauma sessions as part of our mandatory wellness plan.

Twitter
Turkey Transparency Report
December 2022

Data of Reports Received from Users in Turkey

Below are the data of reports including notice and takedowns we received from users in Turkey for violation of personal rights and privacy under Law No. 5651 between June 1, 2022 and November 30, 2022. For us to be able to process some reports for content removal, the claim needs to be specific and strongly supported. Therefore, Twitter needs more information about some reports and asks reporters to provide more information. Additionally, for more information about Twitter's approach to policy development and enforcement philosophy, please visit: https://help.twitter.com/en/rules-and-policies/enforcement-philosophy.

Issue	Volume of Requests	Action Rate %
Abuse	53,462	28.93%
Hateful Conduct	59,524	26.87%
User Impersonation	10,370	16.26%
Brand Impersonation	328	20.12%
Copyright	3,465	50.13%
Incapacitated users	252	0.00%
Deceased Users	394	0.00%
Trademark	401	2.24%
Privacy Policy	51	1.96%
Private Information	70,019	3.98%
Right to Privacy	1,161	17.31%