India Monthly Report under the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021

Published on 2 July, 2021

FACEBOOK

Scope

The following report is published in accordance with Rule 4(1)(d) of the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021. The report contains information on actions taken against violating content on Facebook and Instagram for content created by users in India and proactive detection rates.

This report captures information for the period from 15 May, 2021 to 15 June, 2021.

The next report will be published on 15 July, containing details of user complaints received and action taken. We expect to publish subsequent editions of the report with a lag of 30-45 days after the reporting period to allow sufficient time for data collection and validation. We will continue to bring more transparency to our work and include more information about our efforts in future reports.

Facebook and Instagram policies

We want Facebook and Instagram to be places where people have a voice. To create conditions where everyone feels comfortable expressing themselves, we must also protect their safety, privacy, dignity and authenticity. This is why we have the Facebook Community Standards and Instagram Community Guidelines, which define what is and is not allowed in our community. Facebook and Instagram share content policies. This means if content is considered violating on Facebook, it is also considered violating on Instagram.

Proactive detection and actioning of violating content

The report describes our efforts to remove harmful content from Facebook and Instagram and demonstrate our continued commitment to making Facebook and Instagram safe and inclusive. In this report, we are sharing two metrics which quantify our efforts on removing content and being able to detect them proactively before a user reports them to us. These metrics are our best estimates for actions we take on content created by users who we predict are users in India.

1. Content Actioned: We measure the number of pieces of content (such as posts, photos, videos or comments) we take action on for going against our standards. This metric shows the scale of our enforcement activity. Taking action could include removing a piece of content from Facebook or Instagram or covering photos or videos that may be disturbing to some audiences with a warning.
2. Proactive Rate: This metric shows the percentage of all content or accounts acted on that we found and flagged before users reported them to us. We use this metric as an indicator of how effectively we detect violations. The rate at which we can proactively detect potentially violating content is high for some violations, meaning we find and flag most

content before users do. This is especially true where we have been able to build machine learning technology that automatically identifies content that might violate our standards.

Table 1: Content Actioned and Proactive Rate on Facebook (15 May, 2021-15 June, 2021)

Policy Area	Content Actioned	Proactive Rate
1. Adult Nudity and Sexual Activity	1.8 M	99.6%
2. Bullying and Harassment	118 K	36.7%
3. Dangerous Organizations and Individuals: Organized Hate	75 K	98.8%
4. Dangerous Organizations and Individuals: Terrorist Propaganda	106 K	99.1%
5. Hate Speech	311 K	96.4%
6. Regulated Goods: Drugs	7.7 K	97.9%
7. Regulated Goods: Firearms	2.0 K	89.4%
8. Suicide and Self-Injury	589 K	99.7%
9. Spam	25 M	99.9%
10. Violent and Graphic Content	2.5 M	99.9%

To understand more about each of these policy areas, visit Facebook's Community Standards page.

Table 2: Content Actioned and Proactive Rate on Instagram (15 May, 2021-15 June, 2021)

Policy Area	Content Actioned	Proactive Rate
1. Adult Nudity and Sexual Activity	490 K	96.5%
2. Bullying and Harassment	108 K	43.1%
3. Dangerous Organizations and Individuals: Organized Hate	6.2 K	87.2%
4. Dangerous Organizations and Individuals: Terrorist Propaganda	5.8 K	99.6%
5. Hate Speech	53 K	81.6%
6. Regulated Goods: Drugs	1.1 K	88.6%
7. Regulated Goods: Firearms	0.2 K	87.1%
8. Suicide and Self-Injury	699 K	99.8%
9. Violent and Graphic Content	668 K	99.7%

To understand more about each of these policy areas, visit Instagram's Community Guidelines page.

Note:

1. These metrics are our best estimates of content we act on and of proactive rates based on the creator of the content and predicted country locations for those users.
2. The metrics in this report are based on definitions and caveats as disclosed in Content Actioned and Proactive Rate.
3. Given that such violations are also highly adversarial, country-level data may be less reliable. For example, bad actors may often try to avoid detection by our systems by masking the country they are coming from. While our enforcement systems are global and will try to account for such behaviour, this makes it very difficult to attribute and report the accounts or content by producer country (where the person who posted content was located). Given the global nature of our platforms where content posted in one country may be viewed almost anywhere across the world, other ways to attribute the country of content removed in a technically feasible and repeatable manner, become almost meaningless. So these estimates should be understood as directional best estimates of the metrics.
4. The metric for 'Spam' on Instagram is not available yet. We are working on new methods to measure and report this metric.
5. This report shares metrics in which we have considerable confidence in their accuracy but is not a comprehensive list of all enforcements across Facebook and Instagram. As we develop metrics for new policy areas, we will continue to expand this report, similar to our Community Standards Enforcement Report.