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

Published on 31st March, 2023

Meta

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 for a period of 28 days on (1) actions taken against violating content on Facebook and Instagram for content created by users in India and proactive detection rates, and (2) information on grievances received from users in India via the grievance mechanisms described below. This report captures information for the period from 1st Feb, 2023 to 28th Feb, 2023.

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.

In this report, we have provided a breakdown of proactive detection and content actioned in 13 policy areas on Facebook and 12 policy areas on Instagram.

Table 1: Content Actioned and Proactive Rate on Facebook (1 Feb, 2023 - 28 Feb, 2023)

Policy Area	Content Actioned	Proactive Rate
1. Adult Nudity and Sexual Activity	2.0 M	91.6
2. Bullying and Harassment	226.7 K	71.1
3. Child Endangerment - Nudity and Physical Abuse	182.0 K	86.4
4. Child Endangerment - Sexual Exploitation	174.0 K	98.3
5. Dangerous Organizations and Individuals: Organized Hate	8.4 K	97.6
6. Dangerous Organizations and Individuals: Terrorism	98.6 K	98
7. Hate Speech	74.0 K	69.2
8. Regulated Goods: Drugs	6.4 K	99.2
9. Regulated Goods: Firearms	7.4 K	97.7
10. Suicide and Self-Injury	499.9 K	97.1
11. Spam	20.0 M	99.1
12. Violent and Graphic Content	1.2 M	97.8
13. Violence and Incitement	341.7 K	80.4

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

Table 2: Content Actioned and Proactive Rate on Instagram (1 Feb, 2023 - 28 Feb, 2023)

Policy Area	Content Actioned	Proactive Rate
1. Adult Nudity and Sexual Activity	638.0 K	97.2
2. Bullying and Harassment	427.4 K	93.7
3. Child Endangerment - Nudity and Physical Abuse	28.1 K	97.9
4. Child Endangerment - Sexual Exploitation	357.2 K	99.7
5. Dangerous Organizations and Individuals: Organized Hate	9.7 K	96.8
6. Dangerous Organizations and Individuals: Terrorism	31.9 K	99.2
7. Hate Speech	32.5 K	86
8. Regulated Goods: Drugs	1.9 K	98.3
9. Regulated Goods: Firearms	1.7 K	97.7
10. Suicide and Self-Injury	1.0 M	99.7
11. Violent and Graphic Content	564.0 K	99.2
12. Violence and Incitement	210.8 K	99

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.

Grievances from users in India

This part of the report captures grievances from users in India received via -

- the contact form published in the Help Center regarding violations of Facebook's Terms of Service, Facebook's Community Standards, or questions about accounts;
- the contact form published in the Help Center regarding violations of Instagram's Terms of Use, Instagram's Community Guidelines, or questions about accounts; and
- the common physical mail address published in the Help Center for both Facebook and Instagram.

The above mechanisms are open to all users in India in the respective Help Center pages and are also searchable via the Help Center search bar.

In this report, we report on the following data associated with the grievances from users in India:

1. Total number of reports received, broken down by reasons why they were reported
2. Total number of reports that we responded to
3. Reports where we provided appropriate tools to resolve the issues
4. Reports where further specialized review was required
5. Reports where we took some action against the reported content -
    a. By actioned content, we mean removing the piece of content from Facebook or Instagram, covering photos or videos that may be disturbing to some audiences with a warning, or disabling accounts. When something on Facebook or Instagram is reported to us as violating local law, but doesn't go against our Community Standards, we may restrict the content's availability in the country where it is alleged to be illegal. This includes legal requests such as court orders.
6. Reports may have been reviewed but not actioned for several reasons, including -
    a. The reported content does not violate any of our policies.
    b. The reporter does not provide us enough information to locate the content or account which they are attempting to report.
    c. The reported account or content does not violate our policies.
    d. The reporter is writing to us to provide feedback regarding our services.
    e. The reporter is writing to us regarding a dispute between themselves and a third party which Facebook is not in a position to arbitrate.
    f. The reporter is writing to us regarding content which is not hosted on our platform.

g. The reporter is writing to us regarding content which they wish to see removed, but only providing the link to an entire profile or page, and we are unable to determine which specific content they are referring to.
h. The reporter is requiring assistance from us to access their account.

Facebook

Between 1st and 28th Feb, we received 1,647 reports through the Indian grievance mechanism, and we responded to 100% of these 1,647 reports. The reports were received under the following categories.

Table 3: Reports received

Category	Number of reports
Fake profile that's pretending to be me	129
Content showing me in nudity/partial nudity or in a sexual act	76
Account has been hacked	290
Lost access to a page or group I used to manage	106
Bullying or Harassment	484
Request access to personal data	0
Inappropriate or Abusive Content	385
Report issue with how Facebook is processing my data	0
Content I appear in that I do not want displayed	25
Other issue	152
Total	**1,647**

Of these incoming reports, we provided tools for users to resolve their issues in 585 cases. These include pre-established channels to report content for specific violations, self-remediation flows where they can download their data, avenues to address account hacked issues etc.

Table 4: Reports where appropriate tools were provided

Category	Number of reports
Fake profile that's pretending to be me	28
Content showing me in nudity/partial nudity or in a sexual act	29
Account has been hacked	26
Lost access to a page or group I used to manage	104
Bullying or Harassment	143
Request access to personal data	0
Inappropriate or Abusive Content	83
Report issue with how Facebook is processing my data	0
Content I appear in that I do not want displayed	25
Other issue	147
Total	**585**

Of the other 1,062 reports where specialized review was needed, we reviewed content as per our policies, and we took action on 379 reports in total. The remaining 683 reports were reviewed but may not have been actioned due to the reasons explained above.

Instagram

Between 1st and 28th Feb, we received 14,216 reports through the Indian grievance mechanism, and we responded to 100% of the 14,216 reports. The reports were received under the following categories.

Table 5: Reports received

Category	Number of reports
Content showing me in nudity/partial nudity or in a sexual act	490
Fake profile that's pretending to be me	1,137
Account has been hacked	1,998
Content I appear in, that I do not want displayed	153
Request access to personal data	0
Bullying or Harassment	5,025
Inappropriate or Abusive Content	5,115
Other issue	298
Total	**14,216**

Of these incoming reports, we provided tools for users to resolve their issues in 3,749 cases. These include pre-established channels to report content for specific violations, self-remediation flows where they can download their data, avenues to address account hacked issues etc.

Table 6: Reports where appropriate tools were provided

Category	Number of reports
Fake profile that's pretending to be me	352
Content showing me in nudity/partial nudity or in a sexual act	136
Account has been hacked	53
Content I appear in, that I do not want displayed	142
Request access to personal data	0
Bullying or Harassment	1,410
Inappropriate or Abusive Content	1,378
Other issue	278
Total	**3,749**

Of the other 10,467 reports where specialized review was needed, we reviewed content as per our policies, and we took action on 3,243 reports in total. The remaining 7,224 reports were reviewed but may not have been actioned due to the reasons explained above.

Note:

- Reports in the above categories containing legal processes, including but not limited to court orders, court decisions, statutory declarations, or cease & desist letters, are escalated for legal review.
- Individuals identifying themselves as law enforcement are requested to submit their requests through the Law Enforcement Online Request System (LEORS), an online portal for law enforcement reporting wherein their identity may be verified.
- Reports from a government official or a court officer seeking to submit an order, notice or direction are escalated for legal review.
- In Intellectual Property infringement cases, we direct the reporter to the form on 'Reporting a Violation or Infringement of Your Rights'. The data on grievances received from users in India for this category is excluded from this report.
- In some cases, our reviewers may determine that an issue was mis-categorized in the original report, and escalate the report to specialized review under the right categories. Where applicable, we will include these in the final numbers for actions taken.
- In certain cases, reporters may include multiple URLs in their report. If an action is taken on any one of these URLs in a single report, the report will be listed as actioned.