Internews

Transparency Without Accountability?

FIXING FACEBOOK'S COMMUNITY STANDARDS ENFORCEMENT REPORT

RAFIQ COPELAND, PLATFORM ACCOUNTABILTY ADVISER, INTERNEWS

Introduction

Given that Facebook's Community Standards Enforcement Report (CSER) is considered the industry gold standard, it may seem counterproductive to criticize it. There is an argument that platforms such as YouTube or Twitter will not follow Facebook's lead in transparency if the CSER is overly picked apart. But as Facebook prepares for a new audit of its transparency metrics it feels more important than ever to point out that the CSER provides almost no value as an accountability tool. For outsiders wishing to understand and critically assess Facebook's enforcement of its community standards - which is nominally the idea of this transparency - there is almost nothing in these quarterly updates that would assist with a meaningful analysis. In its current form the industry 'gold standard' amounts to little more than transparency theatre.

To be sure, Facebook provides us with a lot of very big numbers. For example we learn that Facebook 'acted on' 35.7million pieces of content for violating their policies on adult nudity and sexual content in the second quarter of 2020, down from 39.5million pieces of content acted on under the same policies in the previous quarter. In the same period 4million pieces of content were acted on for violating policies relating to 'organized hate', and 8.7million for policies related to terrorism. These numbers are striking - but how should we interpret them? Are they good? Bad? What can they tell us about what is happening in the world, both on Facebook and offline? Do the metrics provided allow us to make an assessment of whether Facebook is doing an adequate job enforcing its rules and keeping its users safe?

The first thing we are likely to notice about the numbers provided by Facebook each quarter is how big they are. 1.5billion fake accounts disabled, 1.4billion pieces of spam. If the intended effect were to underline the Sisyphean task Facebook faces then the report largely achieves its goal (the CSER also includes Instagram statistics, aggregated and accounted for separately). A look at Facebook's earnings report for the same period tells us the platform has 2.7billion monthly active users, a 12% year-on-year growth. Facebook is a world platform. And herein lies the biggest and most obvious problem with Facebook's CSER in its current format. All the figures provided in the CSER are global aggregates. If Facebook's enforcement of its community standards was uniform across its billions of users this may not be a problem. But enforcement is not uniform, and rather than providing transparency the CSER helps to mask this reality.

Break Out CSER Data by Language and Geography

How many pieces of content were acted on for 'organized hate' in Europe, compared to the US? How many in Asia? How many in India, where the company has been slow to name Hindu nationalist groups as 'hate organizations' and Facebook policy officials have been accused of shielding politicians from sanction for inciting hatred? Facebook tells us they acted on 4million pieces of content globally for 'organized hate' in Q2 of 2020 - but there is simply no way to meaningfully assess if this is a good performance without knowing more.

Language draws an even sharper divide. How many pieces of hate speech were acted on in English compared to Spanish? How many in languages such as Bangla, Bahasa Indonesia, or Burmese? The CSER gives us a breakdown of how much hate speech was detected through Facebook's automated processes (94.5% in Q2 of 2020) but doesn't indicate how much of this was in languages like English where the company has good natural language processing and hate speech classifiers. In the vast majority of languages where automation is far less developed Facebook must rely more heavily on user reports, completely changing the dynamic of enforcement. Ample evidence shows us the company is less able to deal with harmful content in these contexts. If the enforcement of community standards is to be meaningful then these distinctions matter. The great majority of Facebook's 2.7billion MAUs are not native English speakers and almost all of its 12% year-on-year growth is from non-English speaking populations.

After sustained pressure from civil society Facebook has released periodic localized updates for one metric specific to one country - hate speech in Myanmar. However these are not disclosed in a systematic or regular form, and metrics for other countries have not been provided. Even these numbers allow us to make a better assessment of Facebook's performance in a country it has treated as a high priority since 2018. For example in Q2 of 2020 Facebook removed 280,000 pieces of content for hate speech violations in Myanmar, up from 51,000 in the previous quarter. We can compare this to the global figures provided in the CSER to see that Myanmar content made up roughly 0.24% of all the hate speech Facebook took action in Q1, rising to roughly 1.24% in Q2. To put this in context, with an estimated 26million users Myanmar makes up just under 1% of Facebook's user base. Notably Facebook tells us that 97.8% of hate speech identified in Myanmar in the most recent quarter was detected through automated systems - which is higher than the global figure. This indicates both that Facebook's Myanmar language AI has improved drastically in the last two years, and that underreporting by users may continue to be an issue.

Breaking out CSER data by language and geography would make these distinctions clear. Facebook should already be collecting this basic data internally. If Facebook is not currently collecting data on enforcement broken down by language and geography as standard then they are failing to do even the most basic oversight of their own enforcement processes. Equally, if they have these numbers for each country and language there is absolutely nothing stopping them from including it in the quarterly CSERs.

Provide Detailed Prevalence Estimates

The next key area where Facebook's CSER must change in order to provide genuine transparency is its presentation of 'prevalence' as a metric. Facebook makes these estimates of prevalence by taking random samples of content and performing a manual review. This metric is used to estimate the overall prevalence of community standards violations on the platform, separate to the enforcement of those standards. This is obviously crucial to assessing Facebook's enforcement performance. The value of Facebook's action on four million pieces of content for 'organized hate' must be weighed completely differently if there are eight million pieces of this content on the platform than if the figure is 200 million. Assessing performance in enforcement without good estimates of prevalence is simply not possible.

Currently the CSER includes estimates of prevalence for some categories of violation and not for others. Categories that do not currently have any measurement for prevalence include 'dangerous organizations, 'hate speech', 'bullying and harassment', and 'spam'. The reasons provided for Facebook's inability to estimate prevalence vary slightly for each. This is the explanation for the category of 'hate speech':

> We cannot estimate this metric right now. Our prevalence measurement is slowly expanding to cover more languages and regions, to account for cultural context and nuances for individual languages. We are still developing a global metric, although our detection and enforcement of hate speech is very broad across the world.

Of course, we would not need a global metric for prevalence if the enforcement numbers were not presented in global aggregate. Understanding prevalence of community standards violations broken down by geography and language is the most important measurement for Facebook to effectively shape its enforcement processes. If the prevalence of spam in Vietnam is much higher than rates of enforcement then the process of automated detection for Vietnamese may need adjustment. If hate speech is a growing problem in Cote D'Ivoire but user reports and automated detection are not surfacing this content for review, then Facebook has a problem.

Facebook says they are expanding their measurements of prevalence to include 'more languages and regions'. Facebook should immediately share the prevalence estimates they currently have. Telling us how many hate speech violations they detected in Myanmar in Q2 is a good step for transparency; telling us how many hate speech violations are estimated to have been *missed* is a necessary step for accountability. As we see the gaps between enforcement and prevalence close over time we will rightly praise Facebook for their improvement. But until we are provided with the information needed to assess such improvement we have no reason to believe that real improvement is occurring.

Facebook should also be transparent about the languages and regions where they don't currently make prevalence estimates and what their timetable is for getting adding these. If we can't assess Facebook's ability to enforce its standards without prevalence estimates, then neither can Facebook. By failing to gather data on prevalence for dangerous content for these languages and regions Facebook is failing its users and putting people in these countries at risk.

The estimates for prevalence metrics that are currently included in the CSER also fall short of what is required for accountability.

Prevalence figures are presented in the CSER as a percentage of content views. For example in Q1 of 2020 (the most recent period available) an estimated 0.07% to 0.08% of global content views showed violations for violent or graphic content. This metric – which measures reach - is certainly very important and should be included for all categories of violation. Reach is a proxy for impact – if a piece of content is viewed a million times it likely has far greater impact that one that is viewed only a handful of times. The problem with *only* reporting prevalence as a percentage of content views, is that enforcement figures for these violations are provided in 'pieces of content' - in this case 25.4million pieces of content acted on for violent or graphic content in the same period.

Content and content views are apples and oranges – we are unable to clearly relate enforcement to prevalence. Without knowing how many pieces of violating content are posted on Facebook in the Quarter we can't possibly make an assessment of the number of pieces of violating content that were removed. If Facebook removed 280,000 pieces of content for hate speech violations in Q2, how many did they miss? A percentage of content views does not answer this question.

The reason prevalence is such a critical metric for accountability is because it allows us to assess how much violating content Facebook is failing to catch. This is particularly important as automated detection improves and gradually lessens reliance on user reporting. Facebook should continue and significantly expand its sharing of prevalence estimates of violating content measured in content views – but they must also share estimates of the actual number of violating posts. With the prevalence figures provided in the CSER the detail that is most important for assessing Facebook's community standards enforcement remains opaque.

Provide Detailed Information on User Reports

Another key failing of the CSER as an accountability tool is that it does not include data on the number of user reports the company receives on its platforms. While the quarterly reports break down the percentages of that content that has been 'acted on' that were flagged through user reports and automation, figures on the reports themselves are not shared. This means we do not know what percentage of user reports resulted in action. Leaving out this data again makes it almost impossible to assess Facebook's performance. Scale is generally cited as the primary operational challenge in moderating a platform such as Facebook. The number of user reports Facebook's moderators are dealing with is a key part of this picture.

Publication of the Myanmar hate speech metrics allows us to get some sense of scale. In Q2 Facebook says that 97.8% of the 280,000 pieces of content it took action on for hate speech in Myanmar were detected through automation. That means that only about 6100 pieces of content in Myanmar were removed by Facebook for hate speech violations as the result of user reports. What we don't know is how many reports were received. How many pieces of content flagged by users in Myanmar were assessed by Facebook's content moderators? And what percentage of these resulted in action? A lack of Myanmar language speaking content moderators has been acknowledged as a previous failing, and Facebook says they have now hired more than 100. Based on the figures provided it would appear that each Myanmar language speaking moderator removed an average of less than one piece of content flagged by a user for hate speech per day (6100 pieces of content, divided by 100 moderators, divided by 90 days equals 0.67 pieces of user reported content a day). Is an overwhelming percentage of content reported by users in Myanmar found to be non-violating? Or is underreporting of hate speech such a serious problem in that country that Facebook is almost entirely reliant on its newly improved Myanmar language hate speech classifier? More information is needed.

One practical reason for the failure to include data on user reports in the CSER may be that the report is currently structured around enforcement categories -- such as 'adult nudity and sexual content' - but content may be removed for a different violation than it was reported for. For example a post may be reported by a user for harassment but taken down for a hate speech violation. Publishing the numbers of user reports received, broken down by geography and language, would avoid this problem as well as provide key insights into where resources and policy changes may be needed. Facebook should share both the number of user reports received in each region and the number of individual pieces of content that were reported. Overreporting and misreporting are likely serious issues in some regions, whilst underreporting may be the larger problem elsewhere.

To give a an example of how user report data can guide product and policy, if it turns out that Facebook is receiving very low numbers of user reports in Ethiopia despite well documented issues with harassment, hate speech and harmful disinformation in that country, then proactive interventions may be warranted. If underreporting is a significant issue then hiring more Amharic and Oromo speaking moderators may not address the core issue as they may not have enough reports to moderate. Improving automated enforcement in these languages is also likely to rely heavily on increasing the number of user reports, as these reports are used to train the automated processes. In this example Facebook could encourage user reporting by localizing reporting features, including in-app guidance and prompts on understanding enforcement of community standards, and prioritizing these languages for development of automation.

Again, data on the number of user reports received and acted on should be easy for Facebook to calculate and there is nothing preventing them from including it in the CSER.

Include Metrics on Speed and Reach

Two further metrics which are necessary to a fair evaluation of Fakebook's performance at enforcing their community standards may be broadly outlined as ‘speed’ and ‘reach’. The exact definition of these metrics could be sliced in different ways, but the reason why they are needed is clear.

Speed. What is the average time between a piece of content first being reported by a user and being ‘acted on’? Or alternatively, how much time between a piece of content first being reported by a user and first being assessed by a human reviewer? For automated detection a sperate metric should be included, with the time between a piece of content being posted on the platform and first being flagged by the system seeming like the most useful formulation. It would also be essential to understand what percentage of content detected by automation is forwarded to a human reviewer and what percentage is removed without human oversight (auto-deletion). Finally, for the content that is detected through automation and forwarded to human review we would want to know if the average times for this process differ from that for user generated reports.

Reach. This is the hardest metric to present in any meaningful way, but also one of the most essential to understanding the impact of community standards enforcement. How many people were exposed to violating content before it was removed or down ranked? Presenting these figures in any aggregate form is unlikely to provide value - the cumulative figure for people exposed to content violations will be far higher than the total number of users. The inclusion of ‘content views’ in the current CSER prevalence metrics points to a way forward, however this doesn’t tell us how many people viewed violating posts before they were removed. Understanding the reach of content that is ultimately acted on provides us with an important way of thinking about improving community standards enforcement. How much does enforcement speed – measured in the ways described above – affect reach? Are posts from accounts with higher reach currently prioritized in moderation queues? If this process was tweaked would this reduce overall exposure to violating content? Some form of meaningful analysis and transparency is surely possible in this category.

Once again, breaking down these metrics by geography and language is essential for effective accountability. When we know how long it takes to respond to violating content in Sinhala or Tamil compared to languages like English we can begin to assess Facebook's investment in enforcing its community standards in Sri Lanka. When we have a benchmark for response times in these languages we can measure improvement, and compare performance against rising or falling numbers of user reports and automated detection to form an informed view of Facebook's progress. With better transparency metrics we can work towards real accountability.

Provide Details on User Appeals

Currently the CSER includes global figures for how many pieces of content that were found to be violations received a user appeal of that decision, and how many pieces of content were restored following such an appeal. In line with the recommendations throughout this document, these metrics should be broken down by language and geography. Just as underreporting of violating content is a more significant problem in countries and communities with lower digital literacy, these users are also less likely to make use of the appeals processes provided by Facebook. With language and context so important to marginal enforcement decisions, these same users may also be at greater risk of enforcement error. Metrics for the speed of appeals should also be added.

Conclusion

In order for the CSER to be a meaningful accountability tool it needs to be far more comprehensive. Facebook is the industry leader on transparency for these metrics - but that doesn't mean we should continue to praise a reporting process that is obviously flawed. If we want other companies to follow Facebook we must push Facebook to continue to lead in this field.

To make the CSER into a true accountability tool, the following changes must be made:

- □ Enforcement metrics for each category of community standard violation included in the CSER must be broken down by country and language.

- □ Regular and detailed prevalence estimates for each category of community standards violation must be provided, broken down by country and language.

- □ In languages and regions where prevalence estimates are not yet available Facebook should provide a timeline for their inclusion.

□ Prevalence estimates should be provided in comparable metrics to the enforcement figures included in the CSER. E.g. if it is estimated that 10% of violations are being actioned and 90% are going undetected, this should be made clear by the metrics provided.

□ The number of reports submitted by users should be included in the quarterly CSER, broken down by country and language.

□ The number of individual pieces of content reported by users should be included in the quarterly CSER, broken down by country and language.

□ The number of pieces of content that are deleted through automation, without human review, should be included in the CSER, broken down by country and language.

□ Speed of enforcement must be included in the CSER, including average time between a piece of content being first reported by a user and action being taken by Facebook, and the average time between a piece of content being posted and being flagged by automated detection methods. Speed metrics must also be broken down by country and language.

□ Reach of standards-violating content and accounts must be calculated and shared in order to assess both the overall impact of such content on societies and Facebook's content moderation effectiveness. Reach metrics must also be broken down by country and language.

□ User appeals should also be broken down by country and language, and a metric should be added for the average time between a user lodging an appeal and the review being completed.

If Facebook does provide more detailed figures they may not be perfect - for example the ability to determine language through automation is itself a major barrier and will mean estimates may need to be used in many cases. Inclusion of a confidence rating and more transparency around how these metrics are calculated will address this problem.

The reason accountability is important for platforms like Facebook is because they are powerful. The processes they put in place and the decisions they make can have enormous impact in the world. Meaningful transparency will help ensure these impacts are positive.