Platforms, Experts, Tools: Specialised Cyber-Activists Network

Diverging responsiveness on reports by trusted flaggers and general users

4th evaluation of the EU Code of Conduct: sCAN project results

The sCAN project participated in the fourth monitoring of IT companies' adherence to the Code of Conduct on Countering Illegal Hate Speech Online organised by the European Commission between 05 November 2018 and 14 December 2018. During this monitoring, sCAN partners reported 762 cases of illegal online hate speech to the IT companies Facebook, Twitter, YouTube, Instagram, Google+, Dailymotion and Jeuxvidéo. The monitored companies took action in 73% of the cases, by either removing (67%) or geo-blocking (6%) the content.

In order to test the reaction of IT companies to notifications by their general user base, notifications were sent anonymously through publicly available channels. In a second step, cases not removed after notification as general users were reported again through reporting channels available only for trusted flaggers/reporters. **Most IT companies reacted more often on notifications sent by trusted reporters than those sent through reporting channels available to their general users.**

Removal rates not only differed between the reporting channels used to send the notifications. The sCAN partners also observed **country specific differences in the reaction to trusted flagger reports**. Facebook removed 100% of trusted flagger cases reported by sCAN partners from Latvia, Germany, France and the Czech Republic, but only 50% of cases reported by the Austrian partner. Twitter and Instagram applied geo-blocking only to trusted flagger notifications by the French partner.

In the Code of Conduct IT companies pledge to "review the majority of valid notifications for removal of illegal hate speech in less than 24 hours and remove or disable access to such content, if necessary." As the time of review of a report is impossible to asses for external organisations, sCAN partners recorded the time when the notified company took action or provided feedback on the notifications.

During the monitoring period, **two monitored IT companies removed the majority of content reported through public channels in less than 24 hours**: Facebook (76%) and YouTube (58%). **YouTube and Instagram took action on more than 50% of the content reported through trusted reporting channels within 24 hours**. YouTube removed 67% and geo-blocked 8% of the content, Instagram removed 50% and geo-blocked 28% in this period. The sCAN partners consider geo-blocking only partly effective, as the content remains online and methods to bypass geo-blocking are widely known in the online community.

Facebook was the only IT company systematically providing feedback to all its users, while the other IT companies provided feedback more often to trusted reporters than to general users. Overall**,** the IT companies provided feedback to 48% of reports through the channels available to general users (46% in less than 24 hours) and to 55% of reports via the trusted reporting channels (45% in less than 24 hours).

All participating **sCAN organisations** have vast experience in monitoring and identifying hate speech online. Against this background, they **consider the overall removal rate not satisfactory**. **Especially the different treatment of reports from general users or organisations having trusted flagger status has to be critically evaluated**. As most CSOs do not have the resources to conduct a continuous monitoring of all Social Media platforms, **engaging users in the fight against illegal hate speech online is essential** to keep users involved and motivated to report illegal content.

SCAN partners agree that while there is an improved effort from social media companies to remove illegal hate speech, **monitoring exercises must continue with the support of European institutions** in order to maintain and improve results. As users disseminating hate content increasingly moved to closed groups or to other platforms not included in the monitoring exercise, and with the changing discourse among far-right users, as they feel confident to significantly gain influence at the European elections in May 2019, it is clear that new challenges and threats are on the horizon. **Continued efforts of monitoring and counter-action are pivotal in ensuring a safe and respectful online space across the EU and beyond.**

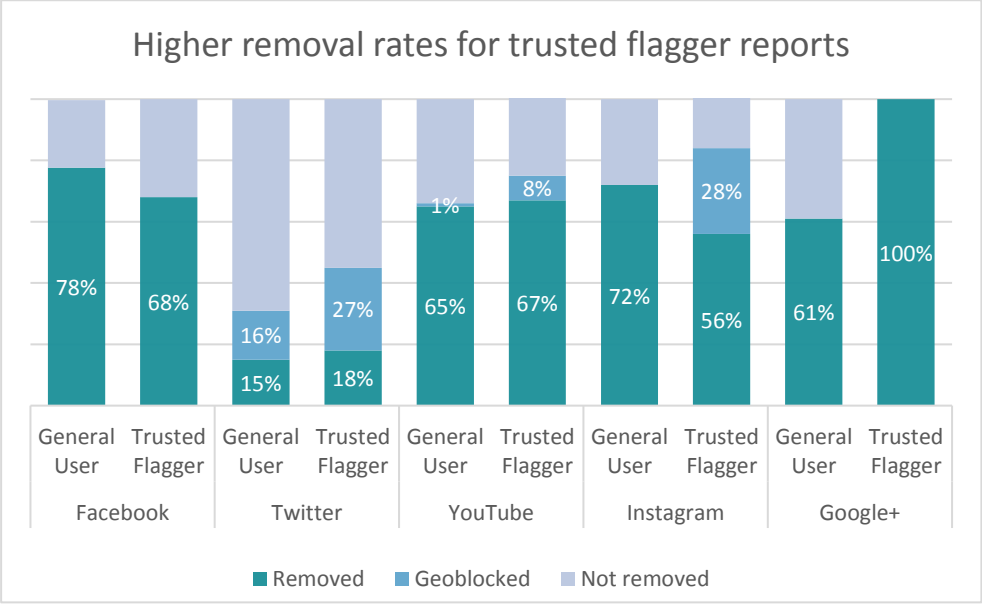
Key Figures:

Between 05.11.2018 and 14.12.2018**,** sCAN partners reported **762 cases of illegal online hate speech** to the IT companies Facebook (311 cases), Twitter (190), YouTube (142), Instagram (86), Google+ (23), Dailymotion (8) and Jeuxvidéo (2). In order to test the reaction of IT companies to notifications by their general user base, **755 notifications** were sent anonymously through **publicly available channels**. In a second step, 165 cases that had not been removed after notification as general users were reported again through reporting channels available only for trusted flaggers. Seven cases were reported directly via the partners' trusted flagger channels. Overall, **172 notifications were sent to the IT companies through the trusted flagger channels**.

The results of this monitoring exercise should not be interpreted as a comprehensive study of the prevalence of hate speech on social media. **They can only provide a momentary picture of content the participating organisations found during a specific six weeks period on the platforms they monitored.** Moreover, the focus of the monitoring exercise was on the reaction of the IT companies rather than the specific content of the illegal hate speech identified.

Removal Rates:

Overall, the monitored companies took action in 73% of cases, by either removing (67%) or geo-blocking (6%) the content. Removal rates differed between the reporting channels used to send the notifications. Most IT companies reacted more often on notifications sent by trusted flaggers than those sent through reporting channels available to general users of the platforms.

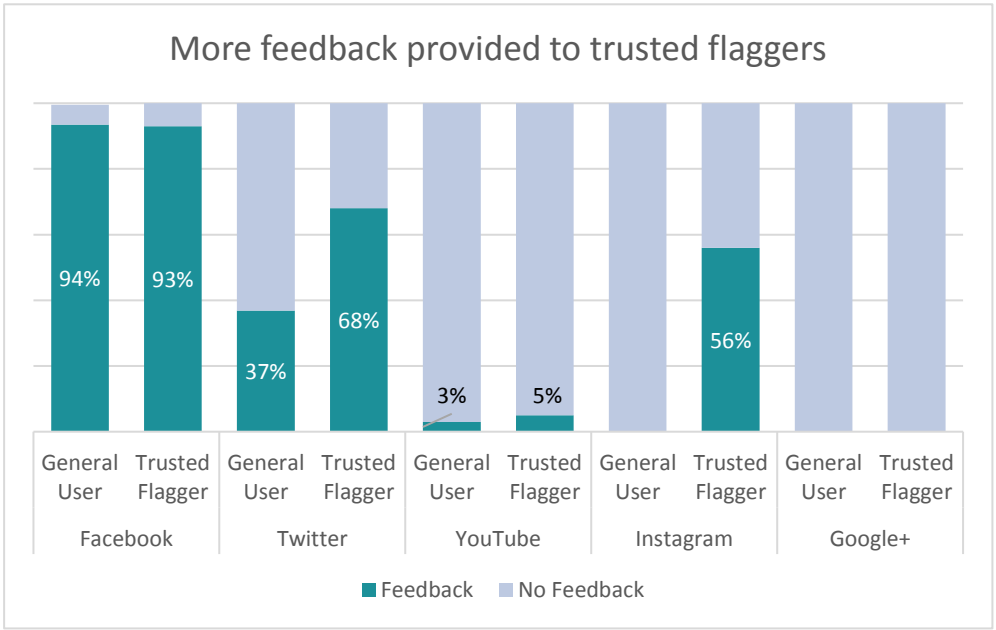


Removal Times:

As the time of review of a report is impossible to asses for external organisations, sCAN partners recorded the time when the notified company took action (removal or geo-blocking). **Two of the monitored IT companies removed the majority of content in less than 24 hours after receiving a notification through the channels available for general users**: Facebook (76%) and YouTube (58%). Instagram removed 47% of this content in less than 24 hours and Google+ 35%. Twitter removed 12% of content within 24 hours and geo-blocked 13%. When reported through trusted flagger channels, YouTube removed the content in 67% and geo-blocked 8% of the cases in less than 24 hours; Instagram removed 50% and geo-blocked 28%, Twitter removed 17% and geo-blocked 27%, while Facebook removed 32% of the content in this period. Google+ removed none of the content reported by trusted flaggers in less than 24 hours.

Feedback:

Overall, the IT companies provided feedback to 48% of reports through the channels available to general users (46% in less than 24 hours) and to 55% of reports via the trusted reporting channels (45% in less than 24 hours). **Facebook was the only IT company systematically providing feedback to all its users**, while Twitter and YouTube provided feedback more often to trusted flaggers than to general users. Instagram provided feedback to trusted flaggers only. Google+ did not provide any feedback during the monitoring period. **Providing feedback on user notifications is essential to keep users involved and motivated to report illegal content to the companies.**



Experiences and observations

During the monitoring period, **Facebook, Twitter and YouTube received the highest number of reports** from each partner organisation. Instagram was included in the monitoring exercise for the first time and some partners had only limited experience in monitoring this platform. However, the partners reported that even though it was more difficult for them to find relevant content on Instagram, they still found an important number of illegal hate speech. Google+ will be closed down in April 2019. Therefore, the partners only sent few reports. Due to their limited reach and relevance only in Croatia and France, Dailymotion and Jeuxvidéo only received a small number of reports.

The well-being of the researchers tasked with conducting the monitoring was an important concern to the sCAN partners, who made sure their staff was well trained and supported throughout the process. In order to ensure privacy and safety for their staff and to mirror the experience of general users while reporting through publicly available reporting channels, partners set up anonymised e-mail addresses and fake profiles on the platforms they monitored.

Several partners observed that **Twitter responded to some reports sent as trusted flagger with a request for more detailed information, including information on the staff member sending the report.** It was not clear to the partners why Twitter would ask for this information. They did not receive any further feedback or assessment after providing the requested information and the content remained online.

Monitoring of the Code of Conduct on countering illegal hate speech online

In 2016, the European Commission and the IT companies Facebook, Twitter, YouTube and Microsoft signed the Code of Conduct on Countering Illegal Hate Speech Online. Google+, Instagram, Snapchat and Dailymotion joined the Code of Conduct in 2018. Between 2016 and 2018 there have been four monitoring periods to evaluate the Code of Conduct. 39 organisations from 26 EU Member States participated in the fourth monitoring from 5 November to 14 December 2018.

You can find an analysis of the results of all participating organisations here: http://europa.eu/rapid/press-release_IP-19-805_en.htm

The sCAN project

Coordinated by the French organization LICRA (International League against Racism and Antisemitism), the sCAN project involves ten different European partners: ZARA – Zivilcourage und Anti-Rassismus-Arbeit from Austria, CEJI-A Jewish contribution to an inclusive Europe from Belgium, Human Rights House Zagreb from Croatia, Romea from Czech Republic, Respect Zone from France, jugendschutz.net from Germany, CESIE from Italy, Latvian Centre For Human Rights from Latvia and the University of Ljubljana, Faculty of Social Sciences from Slovenia. The project aims at gathering expertise, tools, methodology and knowledge on cyber hate and developing transnational comprehensive practices for identifying, analysing, reporting and counteracting online hate speech.

sCAN is funded by the European Commission Directorate – General for Justice and Consumers, within the framework of the Rights, Equality and Citizenship (REC) Programme of the European Union.