SCAN Platforms, Experts, Tools: Specialised Cyber-Activists Network

Monitoring Report

2019 – 2020

Project funded by the European Union's Rights,
Equality and Citizenship Programme (2014-2020)

About the Project

The EU-funded project sCAN – Platforms, Experts, Tools: Specialised Cyber-Activists Network (2018-2020), coordinated by Licra (International League Against Racism and Antisemitism), 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. This project draws on the results of successful European projects already realised, for example the project "Research, Report, Remove: Countering Cyber-Hate phenomena" and "Facing Facts", and strives to continue, emphasize and strengthen the initiatives developed by civil society for counteracting hate speech.

Through cross-European cooperation, the project partners are enhancing and (further) intensifying their fruitful collaboration. The sCAN project partners are contributing to selecting and providing relevant automated monitoring tools to improve the detection of hateful content. Another key aspect of sCAN is the strengthening of the monitoring actions (e.g. the monitoring exercises) set up by the European Commission. The project partners are also jointly gathering knowledge and findings to better identify, ex-plain and understand trends of cyber hate at a transnational level. Furthermore, this project aims to develop cross-European capacity by providing e-learning courses for cyber-activists, moderators and tutors through the Facing Facts Online platform.

sCAN is implemented by ten different European partners, namely 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, Licra – International League Against Racism and Antisemitism from France, 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 sCAN project 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.

Legal Disclaimer

This monitoring report was co-funded by the European Union's Rights, Equality and Citizenship Programme (2014-2020) and the Ministry of Culture of the Republic of Latvia.
The content of the Policy Paper represents the views of the authors only and is the sole responsibility of the sCAN project consortium. The European Commission does not accept any responsibility for use that may be made of the information it contains.

This project is funded by
the European Union

Kultūras ministrija

Content

Introduction

During the second year of the project, the sCAN partner organisations participated in two monitoring exercises, one with the European Commission and one with the International Network Against Cyber Hate (INACH) and the project Open Code for Hate-Free Communication (OpCode). The goal of the monitoring exercises was to evaluate the adherence of the IT companies Facebook, Twitter, YouTube and Instagram to the Code of Conduct on countering illegal hate speech online, developed in 2016 by the European Commission. The sCAN partners have already been participating in previous monitoring exercises organised by the European Commission and INACH.

In the Code of Conduct, the IT companies agree to "review the majority of valid notifications for removal of illegal hate speech in less than 24 hours" and to remove or restrict access to content that violates their Community Guidelines and/or national law. 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.

Between November 4th 2019 and December 13th 2019, the sCAN partners participated in the fifth monitoring exercise organised by the European Commission since 2016. During this monitoring, the partners reported 635 cases of illegal online hate speech to the IT companies Facebook, Twitter, YouTube, Instagram, Dailymotion and Jeuxvidéo.

During January 20th 2020 and February 28th 2020, the sCAN project cooperated in organising an unannounced monitoring with INACH and the project OpCode. The timing of this monitoring was chosen to accommodate the sCAN project duration until the end of April 2020. During this monitoring, the sCAN partners reported 484 cases of illegal online hate speech to the IT companies Facebook, Twitter, YouTube, and Instagram.

Nine sCAN partners contributed to the monitoring exercises:

- ZARA (Austria)
- CEJI (Belgium)
- Human Rights House Zagreb (Croatia)
- Romea (Czech Republic)
- Licra (France)
- jugendschutz.net (Germany)
- CESIE (Italy)
- Latvian Center for Human Rights (Latvia)
- University of Ljubljana, Faculty of Social Sciences (UL-FDV) (Slovenia)

Besides the sCAN organisations, the INACH secretariat and the partner organisations of the project OpCode – ActiveWatch (Romania), DigiQ (Slovakia), Estonian Human Rights Centre (Estonia), Movimiento contra la Intolerancia (Spain) and Never Again (Poland) – participated in the monitoring. For reasons of comparability, the sCAN monitoring report only includes the cases reported by the sCAN project partners.

The results of this monitoring exercise should not be interpreted as a comprehensive study on 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. Some participating organisations focus their work mainly on specific types of online hate speech. This can have an impact on the type of cases reported during the monitoring and will be discussed further below. Furthermore, 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.

Methodology

As in the previous monitoring exercises, the methodology followed closely the monitoring process established by the European Commission during the previous monitoring periods. In a first step, the participating organisations collected instances of illegal hate speech on the social media platforms who joined the EU's Code of Conduct on countering illegal hate speech online. The illegality of the content was assessed against the national laws transposing the Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law[1].

In order to test the IT companies' response to notifications from their general user base, the content was first reported through the public reporting channels of the respective companies. Following this report, the partner organisations recorded whether the IT companies acted on the report by either removing or restricting (geo-blocking, limited features etc.) the content within mutually agreed time periods (24h, 48h, 1 week). Additionally, the partners recorded whether and when they received feedback on their report by the IT companies. Providing feedback on user notifications is essential to keep users involved and motivated to report illegal content to the companies.

Some partner organisations participated in an additional monitoring step by reporting content that was not removed within one week after the initial report via reporting channels available only to organisations recognized by the IT companies as "trusted flaggers". Following this second reporting, the partner organisations again followed the process of the monitoring and recorded the reaction and feedback of the IT companies.

The sCAN organisations agreed to distinguish between content that was removed from the platform and content that was restricted by the IT companies but not removed. Almost all (99%) restricted content was geo-blocked, making it unavailable to users logging in from the country the content was originally reported from. Other forms of restriction include the limiting of certain features (such as comments) on the content or labelling it as sensitive content. The sCAN partners consider restricting content only partly effective, as the content remains online and methods to bypass the restrictions are widely known in the online community.

For the first monitoring exercise covered in this report, the data collection was conducted through an online template designed and managed by the European Commission. The cases were additionally recorded in excel files for internal analysis by the sCAN partners. For the second monitoring, the partners agreed to use a standardised excel template based on suggestions from sCAN partners and prepared by the INACH secretariat.

[1] European Union (2008). *COUNCIL FRAMEWORK DECISION 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law*. Availabel at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008F0913&from=EN (last accessed 26.03.2020).

Key Figures

The results of this monitoring exercise should not be interpreted as a comprehensive study on 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. Some participating organisations focus their work mainly on some types of online hate speech. This can have an impact on the cases reported during the monitoring and will be discussed further below. Furthermore, 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.

Third Monitoring: November 4th to December 13th 2019

The third sCAN monitoring was conducted during the monitoring exercise organised by the European Commission from November 4th to December 13th 2019. During this six week period, the sCAN partners reported 635 instances of illegal hate speech to the IT companies Facebook, Instagram, Twitter, YouTube, Dailymotion and Jeuxvideo. Facebook received the most report from the sCAN partners (280 cases), followed by Twitter with 198 cases. YouTube received 102 reports about illegal hate speech and Instagram received 37 such reports from the sCAN partners.

84 cases were escalated through the channels available only to trusted flaggers of the IT companies, after not having been removed within a week after the initial report through general user reporting channels. Twitter received 59 trusted flagger reports, Facebook and Instagram received 10 trusted flagger reports each and YouTube received 5 reports through trusted flagger channels. No cases were escalated to Dailymotion and Jeuxvideo.

Hate type analysis

The partners recorded the grounds of hatred underlying the illegal hate speech content in accordance with the categories prepared by the European Commission. The most prevalent hate type in the sample was xenophobia, which included hatred against refugees (31% of cases). In the experience of INACH[2] and projects previously conducted by the sCAN partners, anti-refugee hatred emerged as a widespread hate phenomenon in 2015, with the start of the so-called "refugee crisis"[3]. We would therefore urge to separate the numbers for xenophobia not related to the (perceived) refugee status of the target from distinctly anti-refugee hatred in further analyses.

The second most prevalent hate type was anti-Muslim hatred (15 %). Anti-Muslim hatred is often linked with anti-refugee hatred, as haters tend to consider all refugees to be Muslim and all Muslims to be refugees.[4] 11 % of the hate speech cases reported by the sCAN partners were based on racism.

[2] INACH (2016). *"Kick them back into the sea" – Online hate speech against refugees*. Available at https://www.inach.net/kick-them-back-into-the-sea/ (last accessed 26.03.2020).

[3] The term is normatively charged, because it suggests that refugees themselves are problematic or that welcoming refugees per se is critical. However, in our understanding, "refugee crisis" addresses the highly confrontational public debates, the increasing 'scandalization' of migration and the hate filled atmosphere towards refugees.

[4] For more information on hate speech based on an intersection of perceived religion and ethnic origin, see sCAN project (2020). *Intersectional Hate Speech Online*. Available at http://scan-project.eu/wp-content/uploads/sCAN_intersectional_hate_final.pdf (last accessed 26.03.2020).

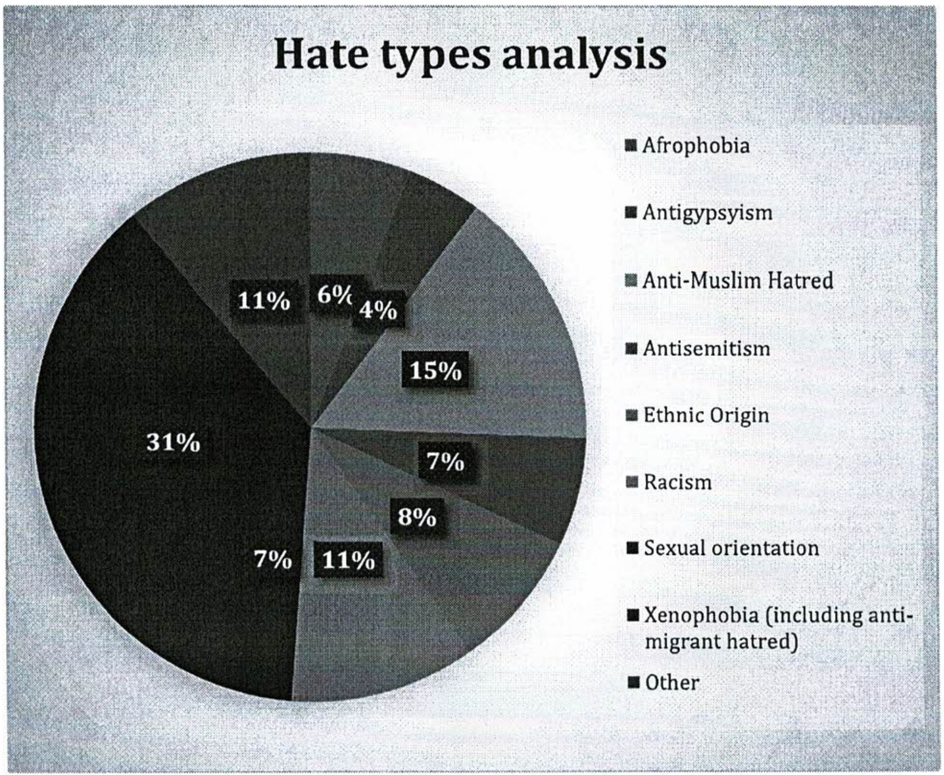

Figure 1: Hate types analysis; sCAN monitoring exercise 4th November – 13th December 2019

Removal Rates

Overall, 67,56 % of the content was no longer available at the end of the monitoring in the country it was reported from (64,25 % removed, 3,31 % restricted). This number is in line with the results of previous monitoring exercises conducted by the sCAN partners. The IT companies acted on 58,74 % of the cases directly after the first reporting through normal user channels (57,80 % removed, 0,94 % restricted). Some partners escalated content that was not removed within a week after initial reporting by reporting it again through the channels available to trusted flaggers. The companies acted on 66,67% of the trusted flagger reports (48,81 % removed, 17,86 % restricted).

Dailymotion and Jeuxvideo had joined the Code of Conduct shortly before the monitoring exercise. Since only very few cases were reported to them (6 and 12 respectively) and they were not included in the fourth monitoring exercise, for reasons of comparability their figures are presented separately. **Jeuxvideo removed 100% of the cases reported to them through general user reporting channels within 24 hours.** Dailymotion removed 33% of the cases reported to them. When they removed cases, they did so within 24 hours. No cases were escalated to Dailymotion and Jeuxvideo.

Of the other monitored platforms, Facebook achieved the highest removal rate (83,21 %) for cases reported through general user reporting channels. YouTube removed 76 % of those cases, Instagram 46 % and Twitter only took action in 16 % of cases by removing 13 % and restricting (geo-blocking) a further 3 %.

All platforms performed considerably better for reports submitted through trusted flagger channels. YouTube removed 100 % of reports submitted by trusted flaggers. Facebook took action on 90 % of cases by restricting 70 % and removing 20 %. It is unclear to the project partners why the platform chose to restrict such a large percentage of cases rather than removing them. Instagram removed 60 % of cases reported by trusted flaggers. The most significant increase in action rate was recorded for Twitter. The company took action on 61 % of cases (47 % removed, 14% restricted), **almost four times as much** as the action taken on cases reported through channels available to their general user base.

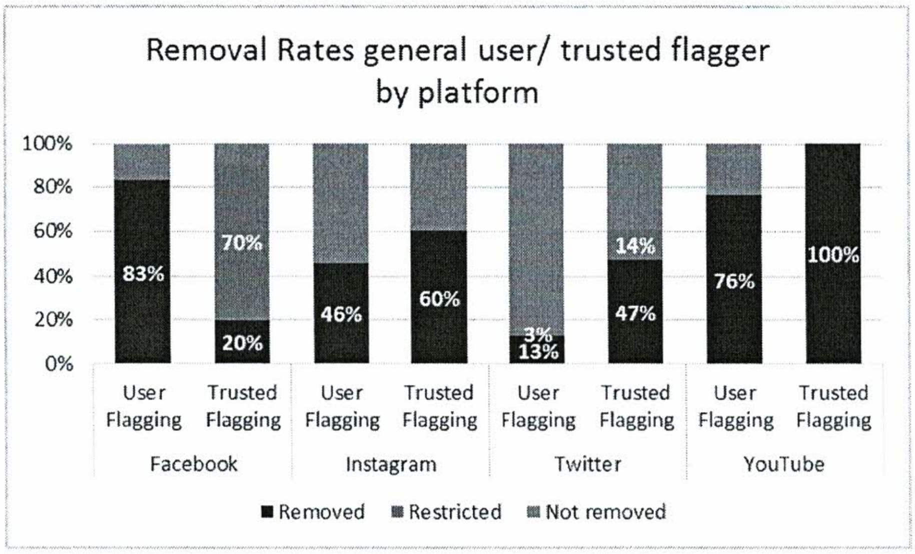

Figure 2: Removal Rates by platform; sCAN monitoring exercise 4th November – 13th December 2019

Removal Times

When it comes to removal times, Facebook again showed the best performance. Facebook was the only platform that took action on the majority of content reported through general user channels within 24 hours (76,43 %). YouTube removed 47,06 % and Instagram removed 40,54 % within 24 hours. Twitter took action on mere 8,59% of content reported through publicly available channels within this time frame.

Most companies reacted more quickly to content reported through trusted flagger reporting channels. Especially Twitter increased its performance to remove 47,46 % of content reported by trusted flaggers within 24 hours. Facebook took action on 70% of this content within 24 hours. Instagram removed 60 % and YouTube removed 40 % of content reported by trusted flaggers within this timeframe.

Feedback

Receiving feedback on reported hate speech on social media is crucial both for users and trusted flaggers in order to build trust and provide users with a better understanding of how the platforms moderate the content. It is also required of the companies by the Code of Conduct signed in 2016 to provide feedback in a timely manner. There is a huge difference between companies involved in the Code of Conduct when it comes to providing feedback. One of the main objectives of the continued organisation of monitoring exercises and of this report is to provide public information on feedback on reported hate speech on social media.

Overall, the IT companies provided feedback to 36,9 % of reports through the channels available to general users and to 56,7 % of reports via the trusted reporting channels. In accordance with previous monitoring exercises, the feedback rate for trusted flaggers is higher compared to the feedback rate for normal users - especially in less than 24 hours (almost 20 percent points of difference). It is important to note that the situation between the IT companies when it comes to feedback rate is very disparate.

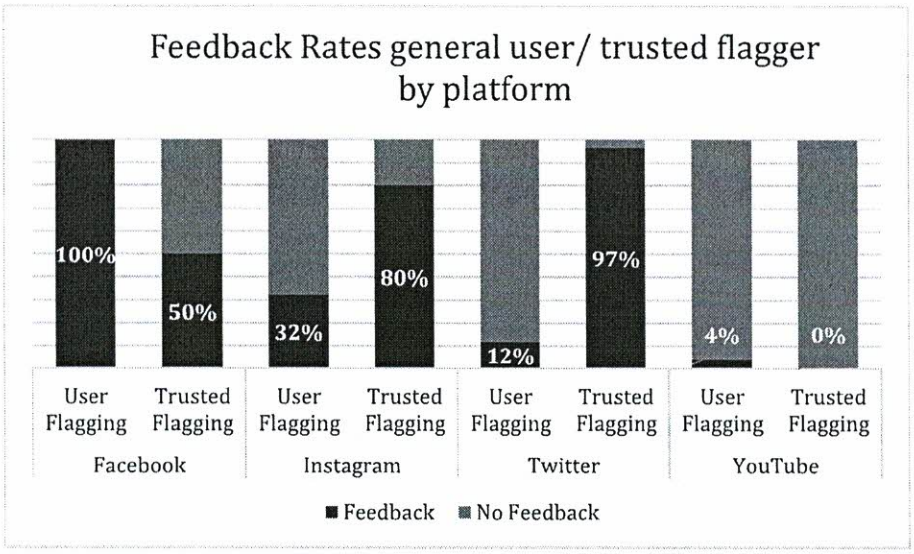

Figure 3: Feedback rates per platform; sCAN monitoring exercise 4th November – 13th December 2019

Facebook may be considered as an isolated case: normal users were almost always provided a feedback after reporting. It was again the only IT company systematically providing feedback to all its users. Nonetheless, trusted flaggers received specific feedback only in 50% of cases.

On Instagram and Twitter, a priority was given to trusted flagger reporting channel regarding their feedback rates (80% and 96,6%). It seems that their policies tend to pay more attention to trusted flaggers. On the other hand, YouTube seldom, if ever, provided feedback to normal users' notifications and trusted flaggers: no feedback has been received for all cases reported by the sCAN partners via trusted flagger reporting channels.

When IT Companies gave feedback, almost all of them provided it within 24 hours. Overall, feedback was provided in 91,2% of cases reported through general user channels in less than 24 hours. Nevertheless this duration could be extended for trusted flaggers: only 75% of feedback was provided in less than 24 hours.

Fourth monitoring: January 20th to February 28th 2020

The fourth sCAN monitoring took place between January 20th 2020 and February 28th 2020. It was a silent monitoring in cooperation with the INACH secretariat and the OpCode project. The sCAN partners reported 484 cases of illegal online hate speech to the IT companies Facebook (242 cases), Twitter (127), YouTube (66) and Instagram (49). In order to test the reaction of the IT companies to notifications by their general user base, the notifications were first sent anonymously through publicly available channels. In a second step, 94 cases that had not been removed after notification as general users were reported again through reporting channels available only for trusted flaggers.

Hate types

To provide an as extensive and in-depth picture of cyber hate within the European Union, the sCAN Project and INACH classified instances of online hate speech during this monitoring exercise into fifteen different categories. An additional 'other' category was also added in order to be able to record cases that would otherwise fall through the cracks. For example, the French partner Licra reported cases of anti-Asian hate speech related to the Covid-19 outbreak, which already received some attention in February 2020.

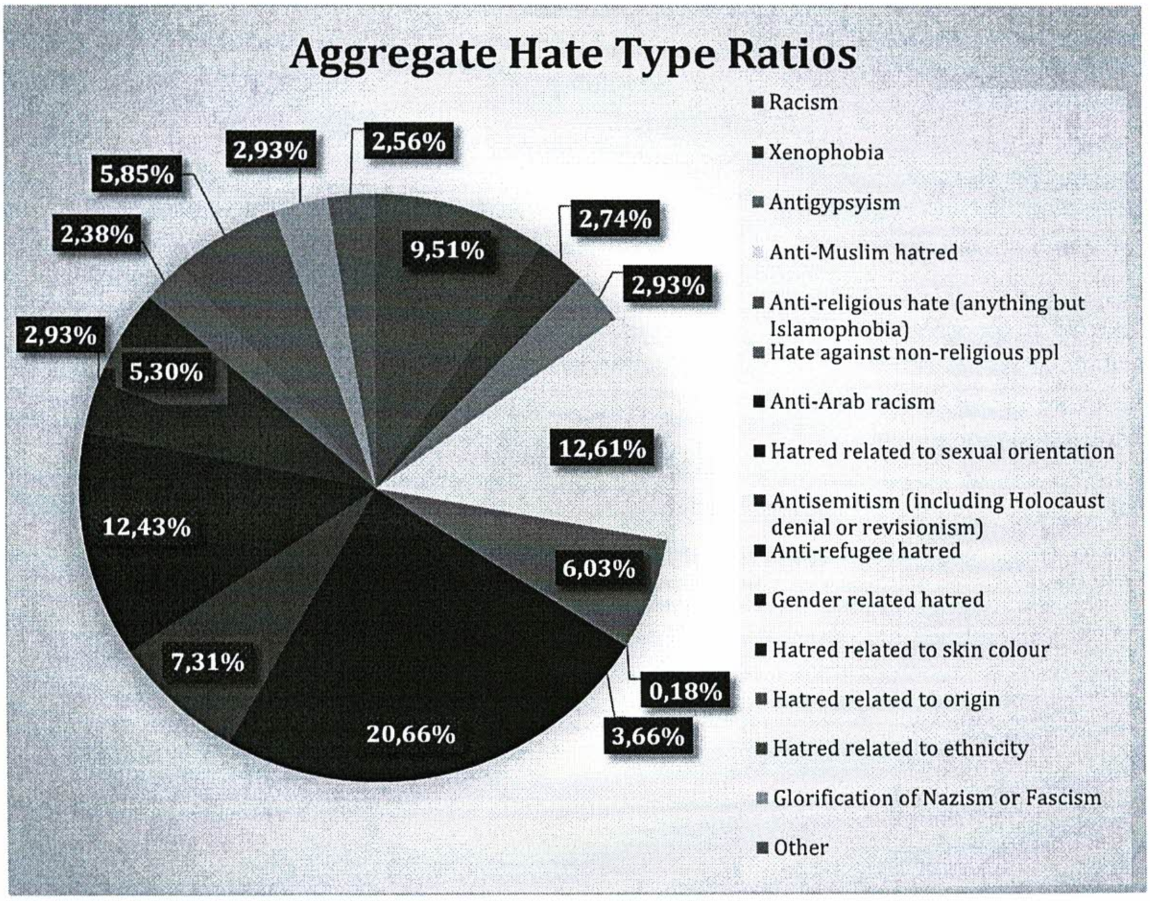

Figure 4: Hate types analysis; sCAN monitoring exercise 20th January – 28th February 2020

Hatred related to sexual orientation was most prevalent hate type in the sample of cases collected during this monitoring exercise. This can be linked to specific, local causes and events that shaped public discussions in some monitored countries during the monitoring period. Most homophobic cases were recorded in Croatia and Latvia. Human Rights House Zagreb (HRHZ) collected altogether 49 cases, out of which 48 were homophobic and the Latvian Centre for Human Rights (LCHR) collected almost 50 homophobic cases, whilst gathering almost double the number of cases than any other organisation participating in the monitoring exercise. Both HRHZ and LCHR reported that during their data collection period there were multiple news stories that pulled LGBTQ+ issues into the centre of public debate. In Croatia, there was an ongoing public debate on whether gay couples would be allowed to be foster parents. In Latvia, there were multiple news stories about a Latvian politician marrying his same sex partner in Germany and a Latvian basketball player having a child with her same sex partner. These events acted as drivers behind hate speech towards the LGBTQ+ community in these countries.

Apart from this particularity, the hate types represented in this sample are consistent with the findings of previous monitoring exercises. Anti-Muslim hatred (12.61%) and anti-refugee hatred (12.43%) top the scale head to head. An unsurprising finding, since the two are intimately interlinked. They are followed by racism (9.51%) and antisemitism (7.31%). Antisemitic hate speech, including holocaust denial, was especially prevalent around the Holocaust Remembrance Day on 27th January.

Removal Rates

Overall, only 58 % of the reported cases were no longer available at the end of the monitoring. This is a major drop compared to the 3rd monitoring exercise conducted only a month earlier. It highlights the importance of a consistent case handling by the platforms, irrespective of official monitoring exercises organised by the European Commission.

51 % of the cases were already removed after the initial notifications as general users (normal user flagging). Instagram achieved the highest removal rate with 75,51 % of cases removed after notification through general user channels. Facebook removed 71,49 % of cases after initial reporting. YouTube and Twitter performed considerably poorer. YouTube removed 25,76 % of cases after user notification, while Twitter only removed 9,45 % and restricted 4,72% of those cases.

94 cases were escalated through trusted flagger channels after not being removed by the companies when reported through general user notification channels. Of those, 39 % were removed by the IT companies. Instagram removed all of the cases reported to them a second time through trusted flagger channels. Facebook removed 68,75 % of the cases reported by trusted flaggers. Twitter removed a considerably higher ratio of cases when they were reported through trusted flagger channels (41,86%) and restricted a further 4,65%, while YouTube removed less cases (6,45 %) than when they were reported by general users.

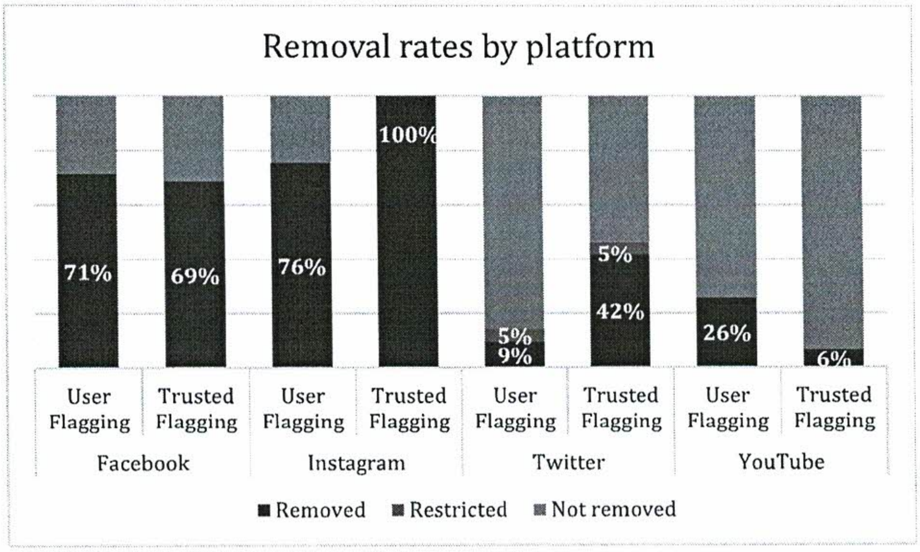

Figure 5: Removal Rates per platform; sCAN monitoring exercise 20th January – 28th February 2020

Assessment and Removal Times

In the Code of Conduct, the platforms commit to assess and remove the majority of illegal cases reported to them in less than 24 hours. The sCAN partners counted the removal of cases and/or provided feedback as assessment. In this monitoring exercise the platforms achieved this goal in 44,21 % of the cases reported as normal users. 4,13 % of cases were assessed after 48 hours and 4,96 % were assessed within a week. In 46,69 % of cases there was no indication of an assessment one week after the initial report.

Only two platforms removed the majority of content reported through general reporting channels within 24 hours. Instagram removed 71,43 % of reported content within 24 hours, its parent company Facebook removed 59 % reported content in this time frame. Twitter (2,36 %) and YouTube (1,52 %) hardly removed any content reported as general users within 24 hours. When the partners reported content through their trusted flagger channels, more content was removed within 24 hours. The biggest difference in removal times between normal user reporting and trusted flagger reporting was reported for Twitter, which removed 37,21 % of content reported by trusted flaggers within 24 hours. Instagram (75 %) and Facebook (62,5 %) further enhanced their performance when content was reported by trusted flaggers. YouTube removed 6,45 % of content reported by trusted flaggers within 24 hours.

Almost three quarters of the cases that were assessed by the companies within a week of the notification were removed by the IT companies, who also provided feedback to the reporting partner. 10 % of the assessed cases were removed, but the reporting organisation did not receive a feedback by the company. In 17 % of the assessed cases the company provided feedback informing the reporting organisation that the content was deemed not against the Community Standards and was therefore not removed.

In the instances when there was no indication of an assessment after a week, the partners checked at the end of the monitoring whether the cases were still online. The vast majority (86,36 %) of those cases were still online after the end of the monitoring and the partners received no feedback on their notification. In 4,09 % of the cases the partners received a feedback more than a week after their reporting to inform them that the content had been removed. In 1,36 % of the cases the content was not removed, but the partners received a feedback notifying them of this decision more than a week after their report to the IT companies. 8,18 % of cases were removed at some point between one week after the notification and the end of the monitoring, but the IT companies did not inform the reporting organisations about the removal. It is therefore impossible to tell if the cases were removed as a result of the monitoring or for different reasons.

The majority of escalated cases were assessed within 24 hours after the report (52 %). 2 % were assessed within 48 hours and 5 % within a week. There was no indication of an assessment in 41 % of the escalated cases. The IT companies did not remove these cases and did not provide feedback to the reporting organisations, even though these organisations are registered as trusted flaggers.

The platforms performed differently when assessing the cases reported to them by trusted flaggers. Twitter (83,72 %), Instagram (75 %) and Facebook (62,5 %) assessed the majority of these cases in less than 24 hours. However, there was no indication of an assessment (either feedback or removal) for cases reported to YouTube as trusted flaggers. Receiving feedback on reported hate speech (even if it is not removed by the platform) is crucial to a fruitful cooperation between the social media platforms and their trusted flaggers. It would be highly welcome, therefore, to receive more communication from YouTube about the reported cases in order to enhance cooperation.

Feedback

The IT companies provided feedback to 51,45 % of reports through the channels available to normal users (42,56 % in less than 24 hours) and to 60 % of reports via the trusted reporting channels (52,63 % in less than 24 hours). According to previous monitoring exercises analysis, the feedback rate for trusted flaggers is higher compared to the feedback rate for normal users - especially in less than 24 hours: almost 10 points of difference for a feedback in less than 24 hours. However, in comparison with the last monitoring exercise, the gap has been reduced by half.

Facebook provided less feedback to users compared to the last monitoring exercise - about 88 %. But, for this exercise, trusted flagger feedbacks rates increased to about 70,6 %.

Instagram's feedback rates considerably improved: user feedback rates almost doubled (from 32 % for the 3rd monitoring exercise to 61 % for this 4th monitoring exercise) and trusted flagger received feedback in 100 % of the cases.

The feedback rates of Twitter and YouTube remained low. Twitter's general user feedback rate has worsened, to 11,6 % for this 4th monitoring exercise. The platform's trusted flagger feedback rate was 96,6 %.

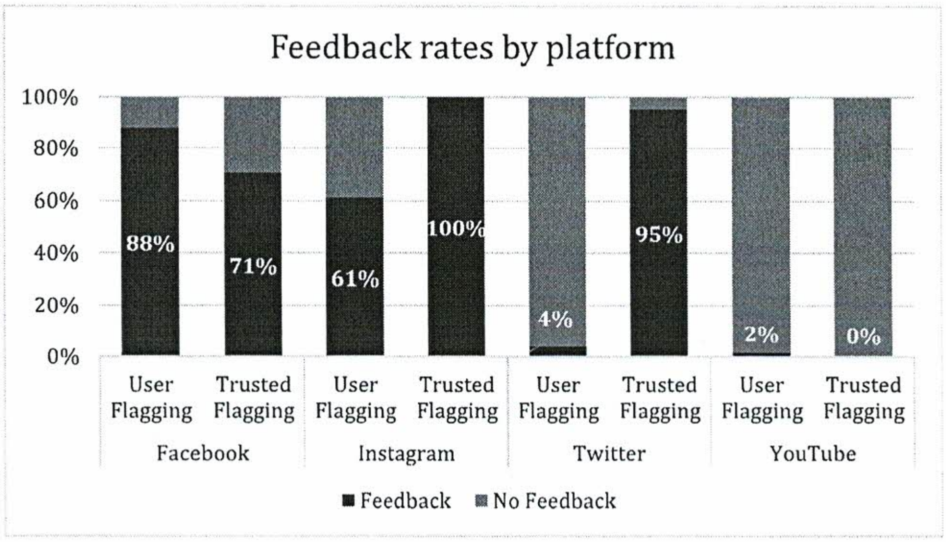

Figure 6: Feedback rates by platform; sCAN monitoring exercise 20th January – 28th February 2020

For YouTube, the situation is basically the same for reports made via normal user channels and through trusted flagger channels and in coherence with the 3rd monitoring exercise: YouTube did not send feedback to any cases reported by the sCAN partners via trusted flagger channels.

Providing feedback took a bit more time to Facebook and Instagram in this 4th monitoring exercise. Facebook provided feedback in less than 24 hours for 71,9 % of cases reported by general user accounts. On Instagram, the rate is 61,2 %.

It is also important to underline that Twitter feedback times have significantly increased: during our 3rd monitoring exercise, in 82 % of cases feedback was provided to normal users within 24 hours. However, this was only true for 1,57 % cases in our 4th monitoring exercise.

Almost half of the platforms sent more feedback to trusted flaggers than to general user reports. For Facebook, as in previous exercises, feedback rate is more important for general user reports compared to trusted flagger reports. Regarding YouTube, as already mentioned, both rate are very low.

Experiences and observations

In order to benefit from the partners' experiences and observations for future monitoring exercises, the sCAN partners filled in an evaluation questionnaire at the end of the monitoring. Some important observations will be discussed in this section.

Partners reported that the device used for reporting seemed to have an impact on whether or not they received feedback from Instagram. While partners reporting through the mobile app reported receiving feedback from the platform, partners reporting to Instagram using a desktop computer hardly received any feedback.

During the monitoring period, partners noticed several accounts posting large amounts of illegal hate speech comments and posts. Some of these pages or accounts have posted a significant number of racist, misogynist and extremely violent comments on a daily basis. Therefore, we recommend that the IT companies monitor those accounts more closely and take decisive action against every instance of illegal hate speech posted on them.

In responding to the question about how the monitoring exercise impacted those who were reporting, the partners emphasized the time consuming impact of the monitoring exercise on their work. Some solved it by distributing the workload among different researchers. There was also an issue of having to deal with increased number of reported cases in the regular stream of work of the organization and therefore decreased capacity to escalate the cases related to the monitoring exercise.

For some organizations with extensive expertise of conducting monitoring exercises, this was just another monitoring and there was no additional impact on the experts, assistants or researchers.

Furthermore, some partners pointed at a range of psychological factors related to the monitoring work. The researcher from ROMEA, for example, reported that the monitoring exercise produced the feelings of "disgust and helplessness" but also determination to carry on.

It is of particular interest that some organizations provide a form of trauma or monitoring-related distress counselling. In case of CESIE, this was done by discussing within the team how the participants' mood was affected by the content they were exposed to. They also reported that one of their researchers became hyperactive in reporting even beyond her working hours and workload. In the case of jugenschutz.net, the employees have access to trauma counselling on a regular basis.

In conclusion, we can say that the monitoring exercises have psychologically and emotionally stressful effects on the reporters, that it is time consuming, has to be preferably shared within the organization by a couple of reporters and that availability of some form of trauma counselling is preferable. Synchronizing the time of the monitoring exercise with the rhythm of the organization can also play a role in the success of the monitoring.

Conclusion

The results of these monitoring exercises highlight the need for a more consistent performance of IT companies in removing illegal hate speech online. The overall removal rate of 58% in the fourth monitoring is almost 10 percentage points lower than the overall removal rate in the previous monitoring exercises. This includes the third sCAN monitoring exercise in November and December 2019, only one month prior. Companies must at all times ensure that they respond in a timely manner and remove illegal online hate speech.

Most companies provide more feedback to trusted flaggers than to their general user base. This can be problematic, as civil society organisations recognized as trusted flaggers cannot monitor and report all illegal hate speech by themselves.

Involving all users of the platforms in reporting hate speech is crucial to combat illegal hate speech online effectively. Feedback is an important aspect to keeping users engaged and motivated to report, as well as to give them a better understanding of how the platforms moderate the content and enforce their community standards.

References

European Union (2008). *COUNCIL FRAMEWORK DECISION 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law*. Availabel at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008F0913&from=EN (last accessed 26.03.2020).

INACH (2016). *"Kick them back into the sea" – Online hate speech against refugees*. Available at https://www.inach.net/kick-them-back-into-the-sea/ (last accessed 26.03.2020).

sCAN project (2020). *Intersectional Hate Speech Online*. Available at http://scan-project.eu/wp-content/uploads/sCAN_intersectional_hate_final.pdf (last accessed 26.03.2020).