Platforms, Experts, Tools: Specialised Cyber-Activists Network

Monitoring Report

2018 – 2019

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 "*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 will enhance and (further) intensify their fruitful collaboration. The **sCAN** project partners will contribute to selecting and providing relevant automated monitoring tools to improve the detection of hateful content. Another key aspect of **sCAN** will be the strengthening of the monitoring actions (e.g. the monitoring exercises) set up by the European Commission. The project partners will also jointly gather knowledge and findings to better identify, explain 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 will be 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, 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 funded by the European Union's Rights, Equality and Citizenship Programme (2014-2020).

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

Content

Introduction

During the first year of the project, the sCAN partner organisations participated in two joint monitoring exercises with the European Commission and the International Network Against Cyber Hate (INACH). 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. Between 2016 and 2018 there have been four monitoring periods to evaluate the Code of Conduct organised by the European Commission. Most sCAN partners have already been participating in the 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"[1] 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.

The first monitoring during the sCAN project duration was organised by the European Commission and conducted 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 second monitoring was jointly organised by the sCAN partners and the International Network Against Cyber Hate (INACH). It was conducted between 6 May 2019 and 21 June 2019. During this monitoring, the partners reported 432 cases to the IT companies 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 INACH partner organisations Greek Helsinki Monitor (Greece) and Never Again Association (Poland) took part in the second monitoring.

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 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.

[1] European Commission (2016). Code of Conduct on countering illegal hate speech online. Available at https://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=54300 (last accessed 22.07.2019).

Methodology

The methodology of the monitoring exercises followed 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 included in the monitoring. The illegality of the content was assessed based on 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[2].

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.

In order to enable the joint analysis and comparison of results, the participating organisations recorded their cases in transnational databases. For the first monitoring, the data collection was conducted through an online template designed and managed by the European Commission. For the second monitoring, the partners agreed to use INACH's database on hate speech. The INACH database was established to provide an international tool to document and analyse instances of cyber hate as well as to function as a central contact point for users to report instances of cyber hate.

[2] 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 22.07.2019).

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.

First Monitoring: November 5th 2018 – December 14th 2018

The first monitoring took place between 05.11.2018 and 14.12.2018. The 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.

Hate type analysis:

For the first monitoring, information on hate types was recorded within the categories set forth by the European Commission for all organisations participating in the monitoring. The most prevalent hate types in the sample of the sCAN project partners were xenophobia (including anti-migrant hate speech) (30%), antisemitism (17%) and anti-Muslim hate speech (12%).

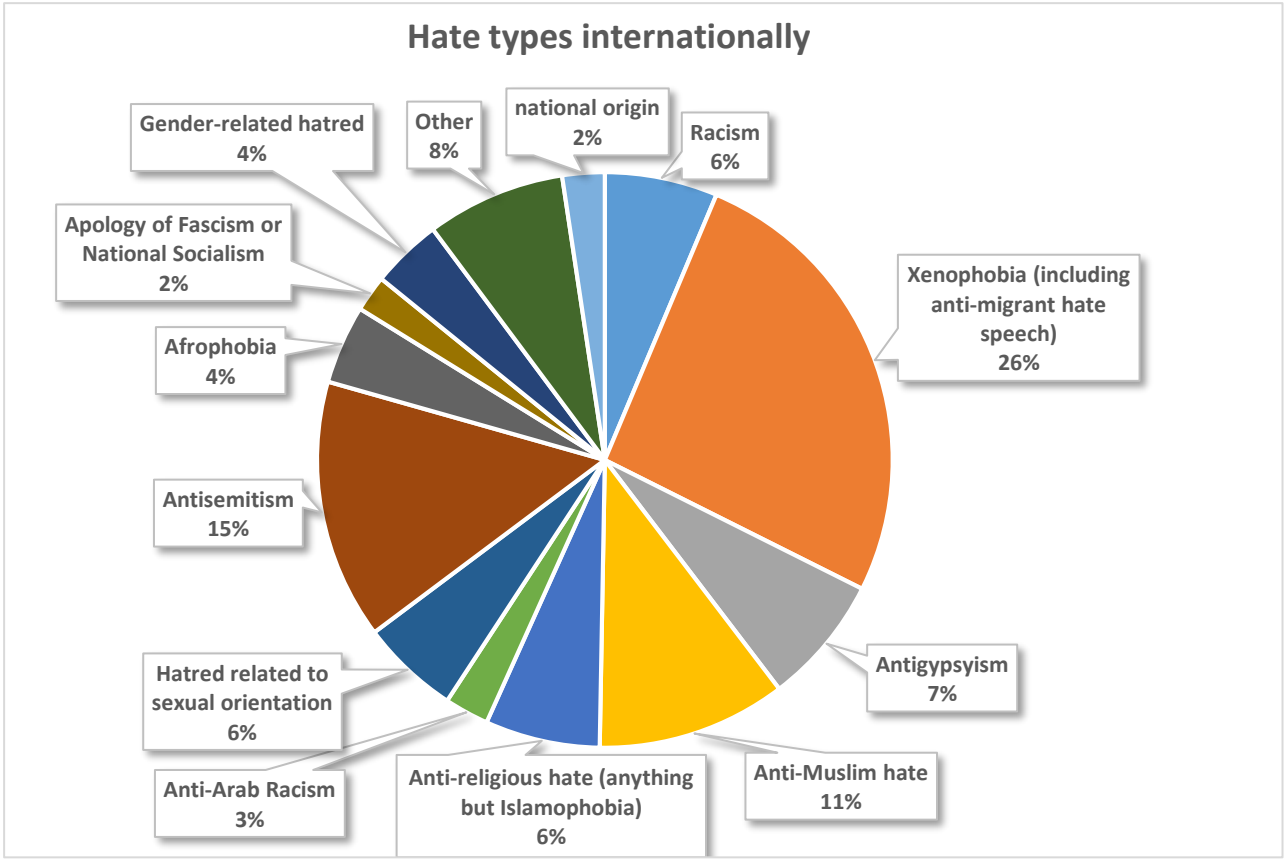

Figure 1: Hate types internationally; Source: sCAN monitoring

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. After reporting through the channels available for general users, the IT companies acted in 62% of cases (58% removed and 4% geo-blocked). When content was reported through trusted flagger channels, the companies acted in 60% of cases (42% removed and 18% geo-blocked). Most companies reacted more often to content reported through trusted flaggers.

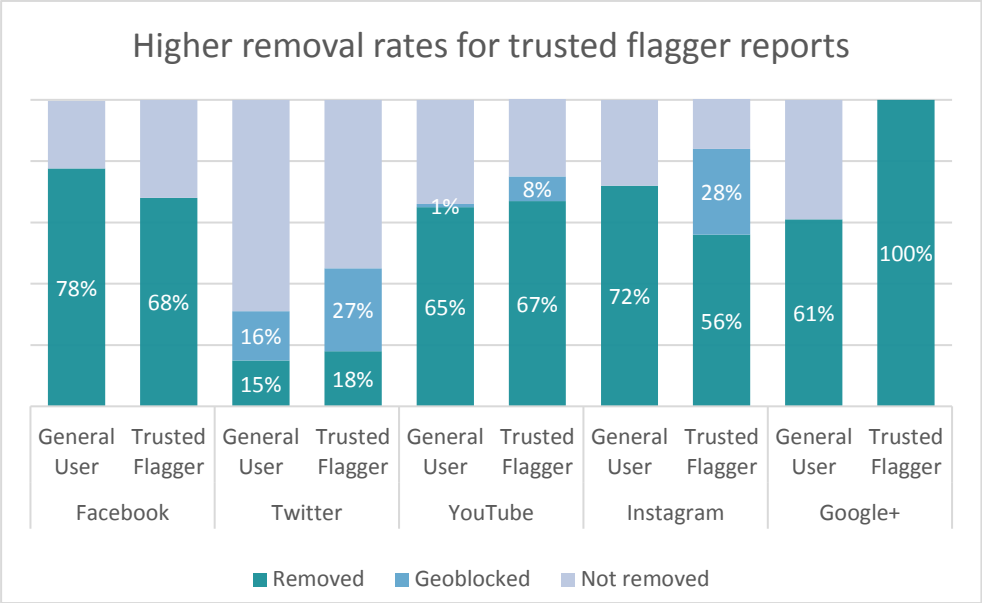

Figure 2: Removal rates general user/ trusted flagger by platform; Source: sCAN monitoring

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 trusted flagger cases reported by the Austrian partner[3]. Twitter and Instagram applied geo-blocking only to trusted flagger notifications by the French partner.

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's removal rate in less than 24 hours was 67%, while 8% of the cases were geo-blocked; 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.

[3] The cases reported as trusted flagger by the Austrian partner, which were not removed by Facebook, were predominantly assessed as relevant to Austrian criminal law by legal experts but display a high degree of complexity. Assessing those cases as illegal and as violation of Facebook's Community Standards requires thorough understanding of relevant context. Assumedly Facebook was not able to grasp the (legal) complexity of the reported cases due to the lack of understanding of national context, knowledge of current political occurrences and a differentiated consideration of a range of German dialects. Facebook did not use the opportunity to consult the trusted flagger.

Feedback:

Overall, the IT companies provided feedback to 48% of reports through the channels available to general users (to 46% of reports 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.

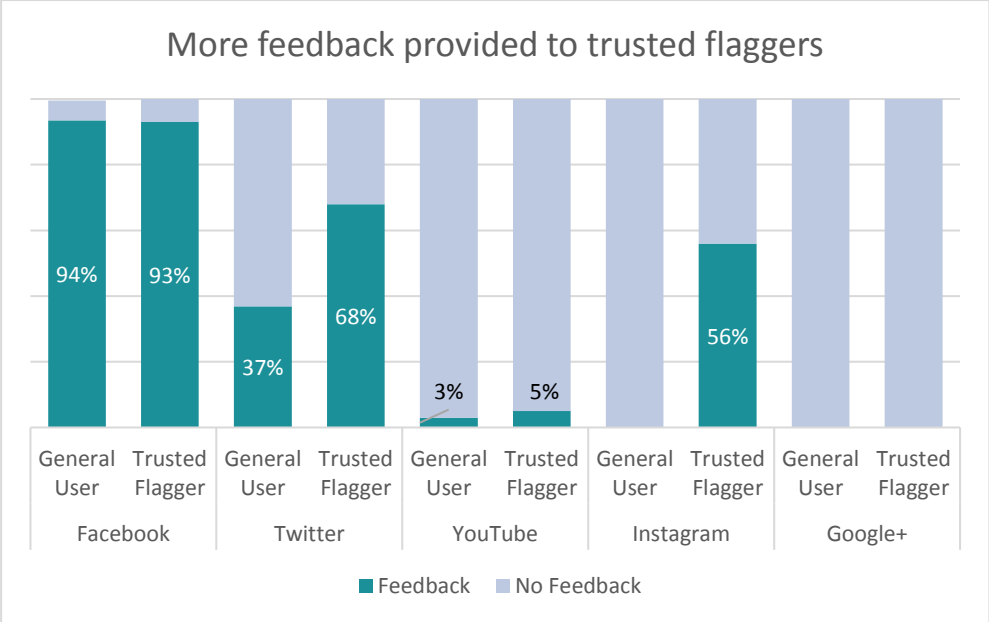

Figure 3: Feedback rate general user/ trusted flagger by platform; Source: sCAN monitoring

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+ 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.

Second Monitoring: May 6th 2019 – June 21st 2019

The second monitoring exercise was implemented between May 6th and June 21st 2019. The sCAN partners reported 432 cases to four different social media platforms. These were: Facebook (200 cases), Twitter (107 cases), YouTube (90 cases) and finally Instagram (35 cases). The IT companies were notified of all these cases through their publicly available channels. After this first round, some sCAN partners reported 90 cases that were not removed through channels that are only available to trusted flaggers.

The partners monitored a plethora of hate types during the exercise. Some of them were more prevalent than others.

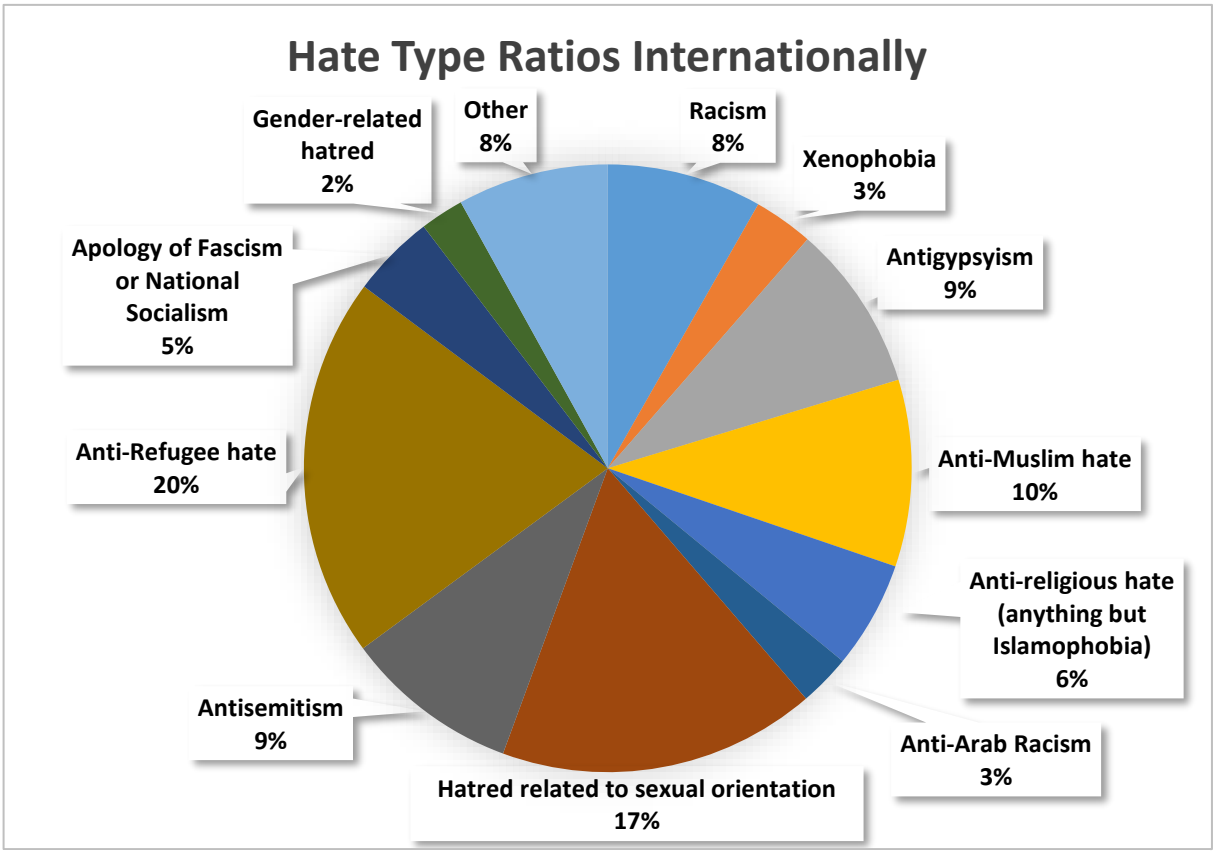

Figure 4: Hate type ratios internationally; Source: sCAN monitoring

The above pie chart gives a snapshot view into the trends in online hate speech. According to our partners' monitoring the most prevalent hate types within the six-week monitoring period were hate against refugees, homophobia, hate against Muslims and antisemitism.

These results can only provide a momentary picture of content the participating organisations found during this specific six weeks period. Some participating organisations focus their work mainly on certain types of online hate speech. In order to better evaluate the hate types found during the monitoring period, the partners provided detailed information about the hate types they reported during the monitoring.

In general, the reported hate types appear to reflect the broader picture of hate speech in the respective countries. In Austria and Slovenia, hate speech against refugees is among the most prevalent according to the experience of the partner organisations. In the Czech Republic, Roma are the minority most consistently targeted with hate speech. Since 2015, hate speech against Muslims, Arabs, refugees and people of colour can be observed more frequently.

In other cases, the hate speech identified during the monitoring was reflecting current discourses and developments within the monitored countries. In Italy, there is a clear incitement to hatred coming from the authority. Most of the hate speech cases reported by the Italian partner were reactions to posts and contents shared by high-ranking politicians or political parties. In France, there has been an antisemitic wave since the beginning of the year. In Croatia, the pride marches held during the monitoring period were targeted particularly by homophobic hate speech online.

In Latvia, antisemitic hate speech was instigated by a draft law about the compensation to the Jewish community for lost communal property during the Holocaust, and by the fact that the newly elected Latvian President is of Latvian and Jewish origin. Homophobic hate speech was triggered by the report about hate crimes against LGBT in Latvia, the draft Co-habitation Law which was turned down by the Parliament and attacks on gay people in Chechnya. Xenophobic hate speech was instigated by discussions about draft law amendments allowing foreign students to work full-time in Latvia.

The German partner, jugendschutz.net, continuously monitors right-wing extremism and Islamist extremism. Islamist cases were mostly targeted at everyone not following Islamist ideology. Most right-wing extremist cases contained glorifications of National Socialism. It is important to note that in Germany the use and dissemination of symbols of unconstitutional organisations is prohibited. The German cases therefore frequently involved symbols associated to Islamist terrorist organisations (e.g. the flag of the so-called IS) or symbols of National Socialist organisations (e.g. swastikas or the SS-skull head).

Removal Rates

Overall, the social media companies removed 67% of content reported during the monitoring and restricted 4%. After reporting through the channels available for general users, the IT companies removed 59% and restricted 3%. Other content was only removed after being reported a second time via the partners' trusted flagger channels. The companies acted on 43% of reports through their trusted flagger channels by removing 40% and restricting 3% of the reported content.

Surprisingly, unlike during previous monitoring exercises, cases reported via trusted flagger channels were not removed in higher ratios on the platforms. Twitter was the only platform where trusted flagging meant a significant and positive difference.

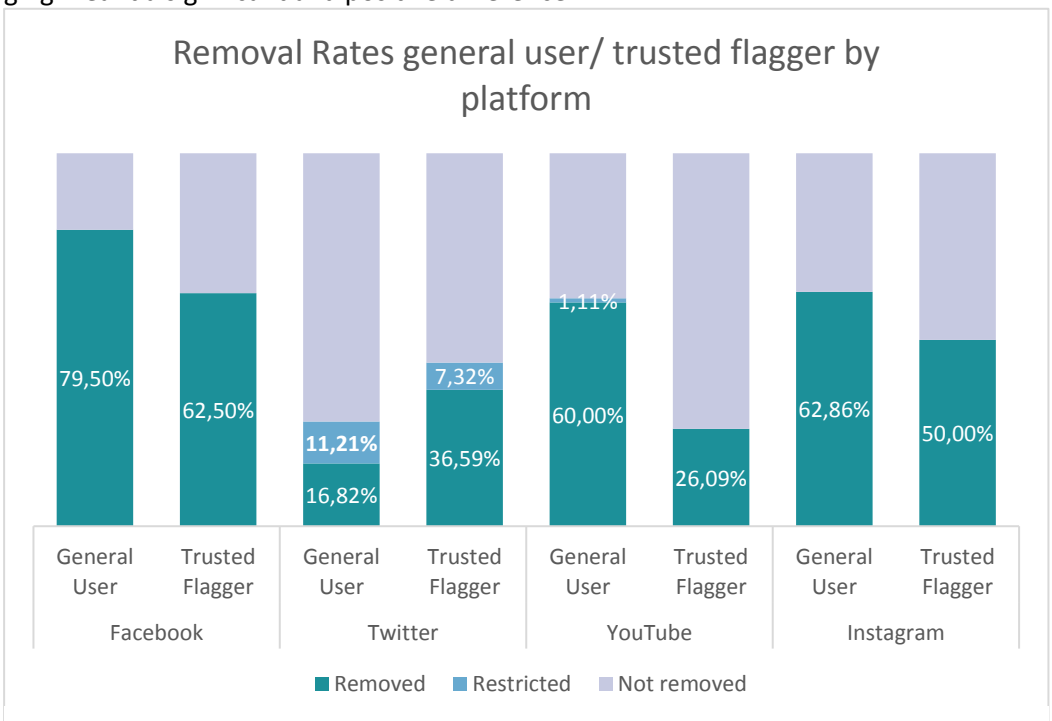

Figure 5: Removal rates general user/ trusted flagger by platform; Source: sCAN monitoring

Nevertheless, it has to be noted that removal rates for cases reported through trusted flagger channels were fairly high, even though these cases had already been rejected once before when reported as a normal user. Thus, it can still be stated that trusted flagging is more effective when it comes to hate speech than the public channels available for all users on the monitored platforms.

Removal Times

The numbers are very varied when it comes to removal rates and times. They vastly depend on the country and on the platform. However, it can be generally said that Facebook is by far the most effective when it comes to removing cyber hate and removing it in a timely manner based on the Code of Conduct.

Of the cases reported through channels available to general users, Facebook removed 64% within 24 hours. Instagram removed or restricted 43% of those cases within 24 hours, YouTube 23%.

Twitter removed or restricted only 17% within this timeframe. Moreover, there were countries where the company did not remove any of the reported hate speech: Romea in Czechia and UL-FDV in Slovenia were not successful in their attempts to have anything removed from the platform. Twitter did not remove anything in Slovenia even when UL-FDV reported as a trusted flagger. In Italy, Twitter only restricted one case reported by CESIE.

The situation is fairly similar when it comes to trusted flagging. Facebook is by far the most responsive company in removing content and removing it within 24 hours (44%). However, Twitter is much better at removing content, especially within 24 hours (37%) when it is reported through the trusted flagger channels. Both Instagram and YouTube had a removal rate of 30% within 24 hours.

Feedback

Receiving feedback on reported hate speech on social media is extremely important both for users and trusted flaggers on social media. It is also required of the companies by the Code of Conduct to provide substantial feedback in a timely manner. There is a huge difference between companies when it comes to providing feedback that is meaningful and timely.

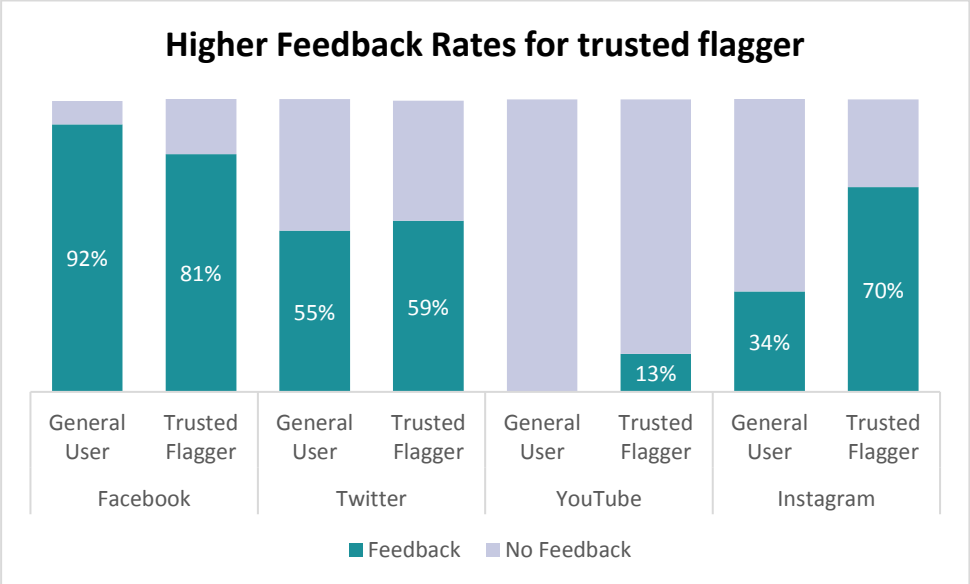

Figure 6: Feedback rates general user/ trusated flagger by platform; Source: sCAN monitoring

Facebook is by far the most efficient in providing feedback. The company responded to 92% of general user reports (74% within 24 hours). Twitter responded to 55% of those notifications (45% within 24 hours), and Instagram responded to 34% of notifications by general users within this timeframe.

YouTube did not send feedback to any cases reported by the sCAN partners via general reporting channels.

As it can be seen on the chart above, most platforms sent more feedback to trusted flaggers than to general user reports. The exception is Facebook that responded more often to general user reports than to trusted flaggers.

Experiences and observations

In order to get a better insight into the partner organisations' experiences during the monitoring, an evaluation questionnaire was disseminated upon conclusion of the monitoring exercise.

The partners reported that only few cases were restricted rather than removed by the IT companies. Of those, the vast majority was geo-blocked. However, the French partner reported that during the second monitoring all cases of homophobic hate speech they reported to Twitter were restricted rather than removed. Since the content thus remains online and methods to bypass the restrictions are widely known in the online community, the sCAN partners consider this approach only partly effective. There was no indication why Twitter would apply it specifically to French homophobic content.

The Czech partner observed that the Czech Republic cannot be chosen as a location in the "trends" section on Twitter. This complicates the monitoring process, as it makes it more difficult to monitor relevant national debates for hateful content. Another problem encountered during the monitoring was that direct links to reported comments on Facebook did not work reliably. In long conversations with a multitude of comments, this makes it very difficult to check if the reported comment was removed.

When providing feedback, the IT companies responded mostly with automated messages not giving details about the specific case or the reasoning behind their decision. Furthermore, some IT companies treated some partners differently than the others. While Facebook had the highest feedback rate and provided feedback to both general users and trusted flaggers, some partners observed that the feedback was not sent immediately when the company took action, but only a few days later.

The Italian partner received customized feedback for all content they reported to Facebook, whereas the Slovenian partner received only automated responses when reporting as general user. For trusted flagger reports, they received feedback via e-mail. However, those e-mails did not always reference the reported content, making the follow-up the cases very complicated.

Conclusion

A comparison between the results of the first and second monitoring shows similar removal rates. However, since Facebook received the most reports from the participating organisations, this does not reflect the performance of all monitored IT companies. Looking at the companies' performance separately, we note that with the exception of Facebook the platforms did not perform as well in the second monitoring. For what concerns the content reported through trusted flagger channels, less cases were removed or restricted by the IT companies than during previous monitoring exercises organised by the European Commission.

In the Code of Conduct, the companies agreed to assess and remove content that is against national law or their Terms of Services within 24 hours. Yet, only Facebook and Instagram managed to reach a tolerable level in removing reported hate speech within that timeframe, while Twitter and YouTube did not reach 50%.

The companies also agreed to provide substantial and timely feedback to reporters of illegal content. The removal of hate speech is important, but feedback to reports is just as important if not more. Providing feedback, even as an automated response, is seen as vital in providing transparency about

their actions towards their users and encouraging them to support social media in combatting hate speech online.

The partners observed a decline in feedback compared to previous monitoring exercises. Facebook is the only company that systematically provided feedback to both general user reports and trusted flagger reports during both monitoring exercises. Facebook was also the only company providing the majority of feedback within 24 hours of reporting. YouTube was specifically criticised for providing hardly any feedback to both normal users and trusted flaggers.

In order to fully implement the Code of Conduct and effectively combat illegal hate speech online, it is crucial that social media platforms react to all reports they receive from their user base in a timely manner, regardless of who is reporting or the reporting period. Continued efforts of monitoring and counter-action are pivotal in ensuring a safe and respectful online space across the EU and beyond.