INACH
bringing the online in line with human rights

SCAN

Monitoring Report

06. May 2019 – 21. June 2019

International Network Against Cyber Hate - INACH

Founded in 2002 by jugendschutz.net, Germany, and Magenta Foundation in the Netherlands, the International Network Against Cyber Hate (INACH) currently unites 29 organizations from Europe, Israel, Russia, South America and the United States. While starting as a network of online complaints offices, INACH today pursues a multi-dimensional approach of intervention and preventive strategies. The member organisations are united in a systematic fight against cyber hate, for example as complaints offices, monitoring offices or online help desks. In their respective countries, they provide important contacts for politicians, internet providers, educational institutions and users.

Funding for INACH is provided by:

European Union Rights, Equality and Citizenship Programme (2014-2020)

Funded by

Federal Ministry for Family Affairs, Senior Citizens, Women and Youth

as part of the federal programme

Demokratie leben!

bpb: Federal Agency for Civic Education

Project sCAN:

Coordinated by the LICRA (International League against Racism and Antisemitism), France, the sCAN project involves ten different European partners: ZARA – Zivilcourage und Anti-Rassismus-Arbeit, Austria, CEJI - A Jewish contribution to an inclusive Europe, Belgium, Human Rights House Zagreb, Croatia, ROMEA, Czech Republic, Respect Zone,France, jugendschutz.net, Germany, CESIE,Italy, Latvian Centre For Human Rights,Latvia and the University of Ljubljana, Faculty of Social Sciences, 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.

Funding for sCAN is provided by:

European Commission - Directorate General for Justice and Consumers, within the framework of the Rights, Equality and Citizenship Programme (2014-2020)

Content

Introduction

Between 6 May 2019 and 21 June 2019, the International Network Against Cyber Hate (INACH) and the sCAN project jointly organised a monitoring exercise 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 by the European Commission in 2016. Between 2016 and 2018 there have been four monitoring periods to evaluate the Code of Conduct organised by the European Commission. Most INACH and sCAN partners have already been participating in the previous monitoring exercises organised by the European Commission and INACH. During the latest monitoring, the participating organisations reported 432 cases to the IT companies Facebook, Twitter, YouTube and Instagram.

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, the partner organisations recorded the time when the notified company took action or provided feedback on the notifications.

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)

Apart from the sCAN organisations, INACH secretariat participated as a coordinator and facilitator with its online database and by inviting two other INACH members to the exercise, the Greek Helsinki Monitor (Greece) and the Never Again Association (Poland) that took part in this monitoring.

Some participating organisations focus their work mainly on specific 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"[3]. 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 organisations participating in the monitoring agreed to distinguish between content that was removed from the platform and content that was restricted by the IT companies but not removed. The majority of 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 partners agreed to use INACH's database on hate speech to record their monitoring cases. 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.

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.

---

[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).

[3] A "trusted flagger status" is a special status granted by social media companies to organisations and institutions that have proven expertise in evaluation and classifying online hate. The status provides for direct communications and channels to the companies' respective departments.

Key Figures

Overall, the partners sent 522 notifications to the IT companies[4]. 432 notifications were sent through publicly available channels. Additionally, 90 notifications were sent through reporting channels available to the partner organisations as trusted flaggers/reporters, after the content was not removed within a week of the general user notification.

Facebook received 200 general user notifications and 16 trusted flagger notifications. Twitter received 107 and 41 notifications respectively. YouTube received 90 notifications through general user channels and 23 through trusted flagger channels. Instagram received 35 and 10 notifications respectively.

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

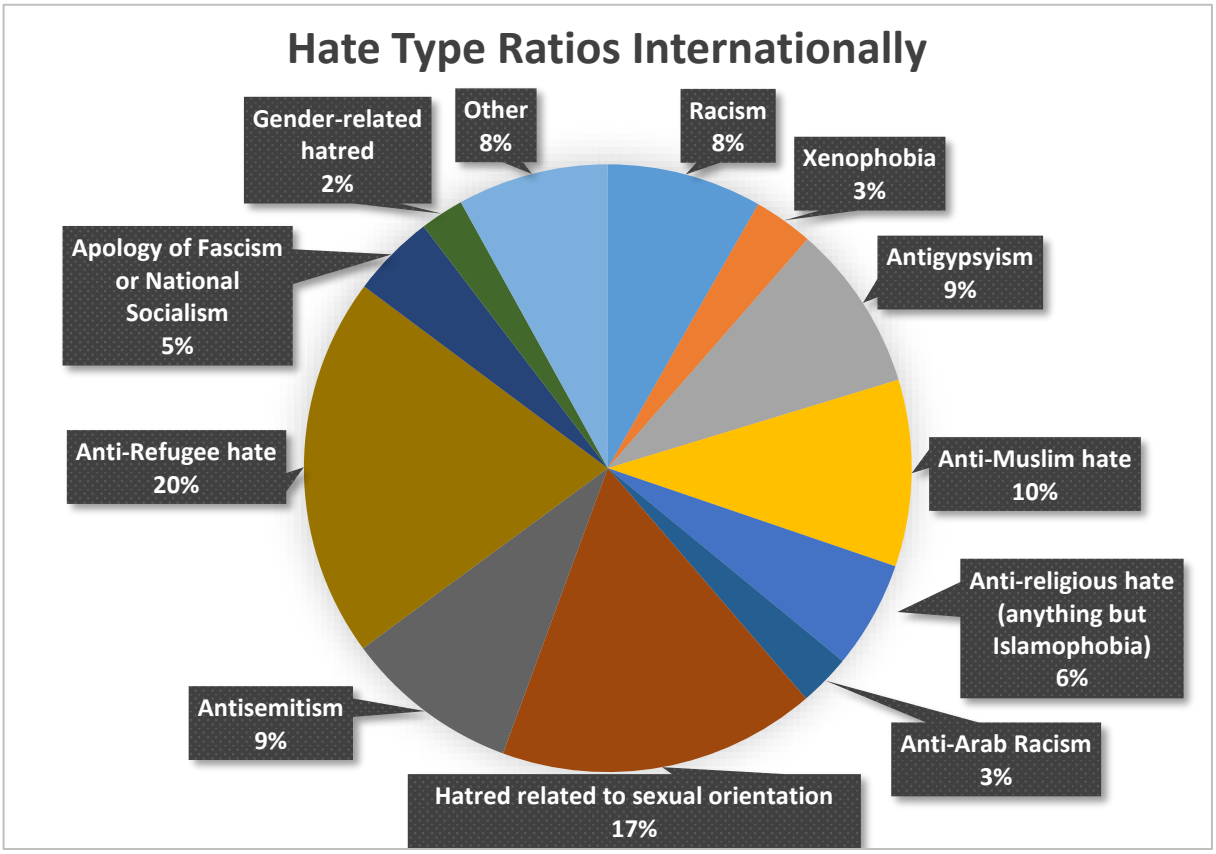

Figure 1: Hate type rations internationally; Source: Joint Monitoring INACH/sCAN project

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

[4] The cases and numbers given are not representative of the prevalence and types of illegal hate speech online in absolute terms, and are based on the number of notifications and cases submitted by the participating organisations.

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. When it comes to the cases actively monitored throughout the six-week monitoring phase, cases of hate speech against refugees were most prevalent in Austria and Slovenia.. 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 Jew-ish 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 insti-gated 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 organisa-tions (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%. Although the overall action rate of 70,6 % turned out to be only slightly lower compared to the last monitoring (-1,1 percentage points, this result is mostly owed to Facebook's consistently high removal rate of 84,5% (+0,9 percentage points) and Instagram's improvement to 77,2% (+6,6 percentage points). Twitter's performance remained low at 44,9% (+1,4 percentage points), and YouTube only removed or restricted 67,8% of illegal hate content, a major drop of 17,6 percent points compared to its last checked performance.

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.

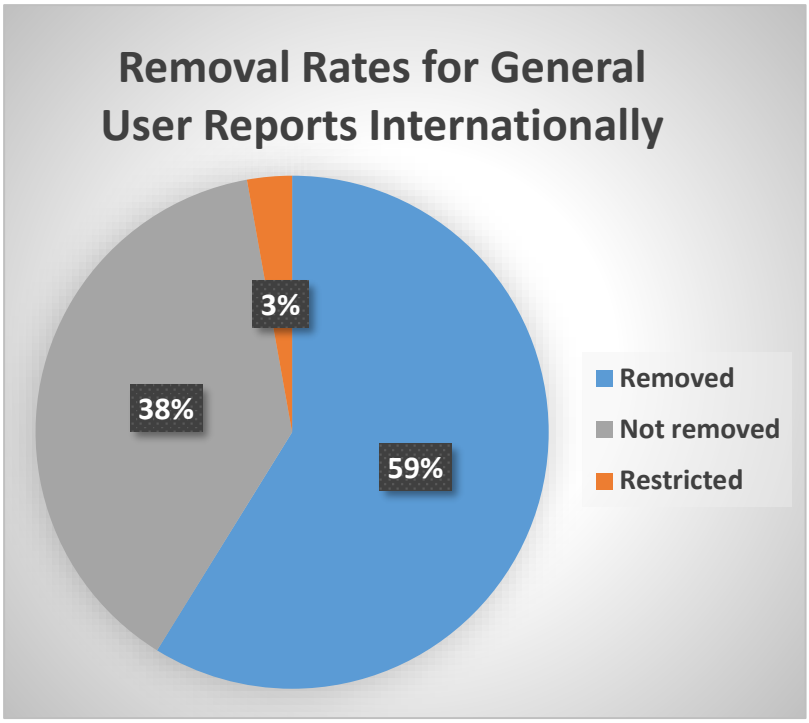

Figure 3: Removal Rates for General User Reports internationally; Source: Joint Monitoring INACH/sCAN project

Surprisingly, unlike during previous monitoring exercises, reporting as trusted flaggers did not result in higher removal rates on all platforms. Twitter was the only platform where the removal rate for trusted flagging was significantly higher than the removal rate for general user flagging (44% compared to 28%).

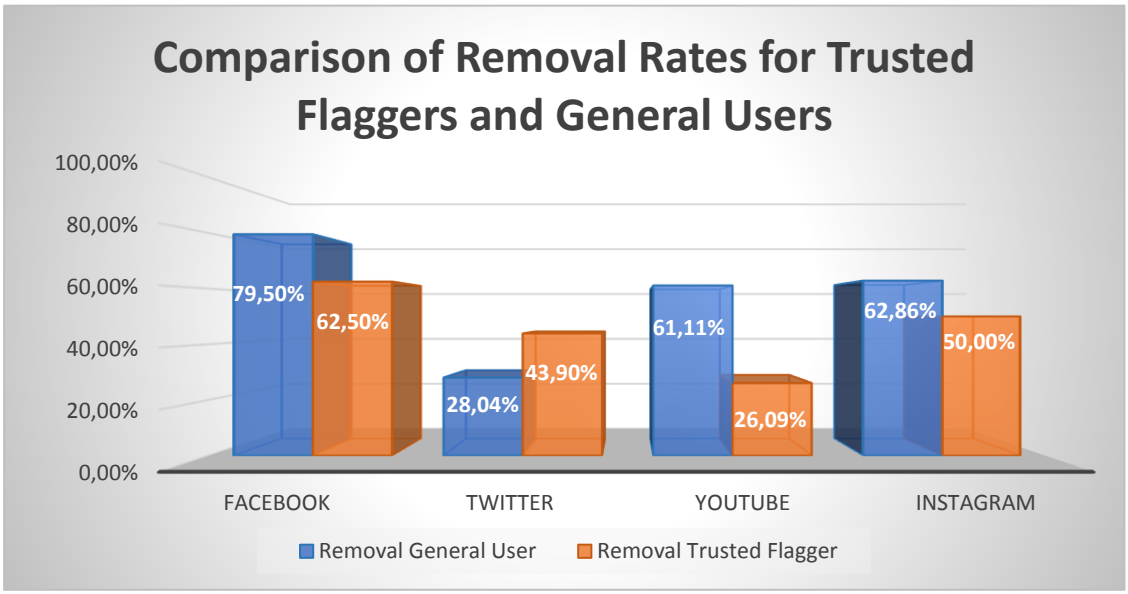

Figure 2: Comparison of Removal Rates for Trusted Flaggers and General Users; Source: Joint Monitoring INACH/sCAN project

Nevertheless, it has to be noted that removal rates for cases reported through trusted flagger channels were fairly high, 43%, even though these cases had already been rejected once before when reported as a general user. Thus, it can be still stated that trusted flagging is still much more effective when it comes to hate speech than the public channels available for all users on the monitored platforms.

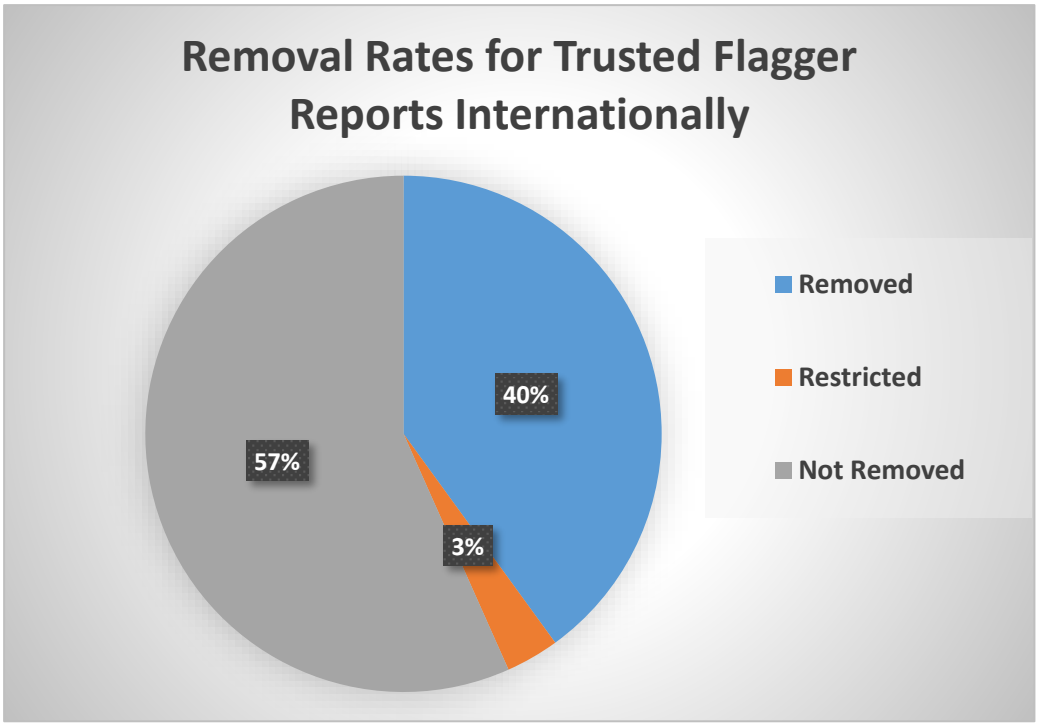

Figure 4: Removal Rates for Trusted Flagger Reports internationally; Source: Joint Monitoring INACH/sCAN project

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, globally it can be 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. The company removed almost 80% of the instances of cyber hate reported as general users and they removed 79,9% of it within 24 hours. YouTube moves in the mediocre range with removing and restricting a bit more than 60% of reported content and only removing/restricting 38% of it within 24 hours. An abysmal outcome from a company that has been participating in the Monitoring processes for years now. Instagram, when it comes to its removal behavior, is located in the middle field, but its reaction by removing hateful content is more satisfactory than YouTube's removal behavior. Still, Facebook has a lot to improve as the parent company of the platform. Twitter exhibited the worst performance during the monitoring exercise. The company hardly removed or restricted any reported content (28%) and removing/restricting 60% of it within 24 hours.

Moreover, there were countries where the company did not remove any of the reported hate speech: Romea in Czechia and UL-FDV in Slovenia did not manage to get anything removed from the platform. Twitter did not remove anything in Slovenia even when content was reported by a trusted flagger. In Italy, Twitter only restricted one case reported by CESIE.

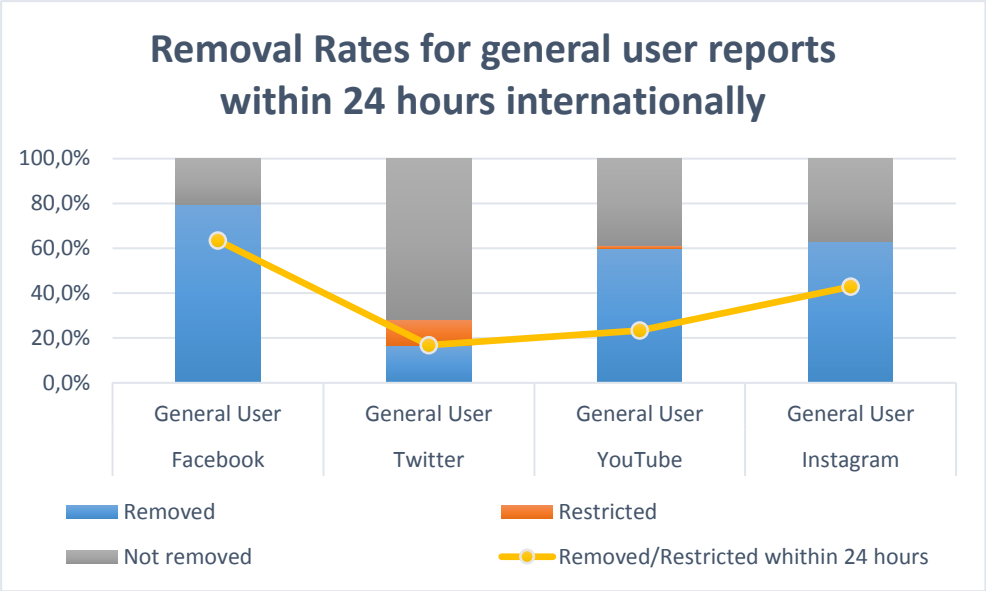

Figure 6: Removal Rates for general user reports within 24 hours internationally; Source: Joint Monitoring INACH/sCAN project

The situation is fairly similar when it comes to trusted flagging. Facebook shows by far the most adequate reaction behaviour when it comes to removing content and removing it within 24 hours. However, Twitter is much better at removing content, especially within 24 hours when it is reported through the trusted flagger channels. Instagram is pretty similar, and YouTube is a bit worse in removing the reported hate speech, but a lot better at removing it within 24 hours. However, it should not be forgotten that these cases had been reported once before as general users, were rejected by the companies by stating that they do not violate their community standards and then escalated through trusted flagger channels. Hence, it is unsurprising that the removal rates are lower. This is due to interpretation differences between the companies and the NGOs when it comes to national law and the platforms' terms of services.

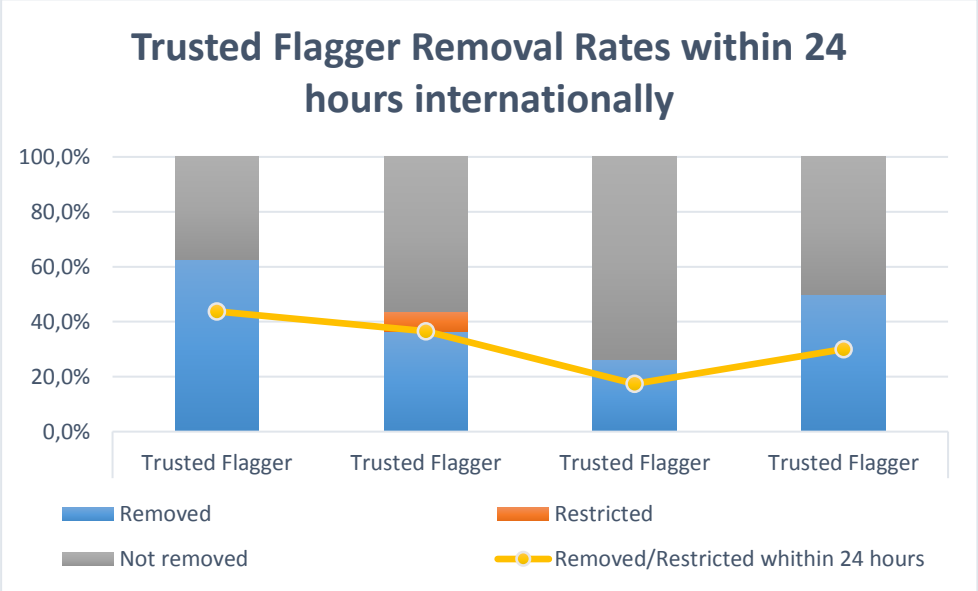

Figure 5: Trusted Flagger Removal Rates within 24 hours internationally; Source: Joint Monitoring INACH/sCAN project

Feedback

Receiving feedback on reported hate speech is extremely important for both general 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. Yet, companies are notoriously bad at providing clear and timely feedback. Naturally, there is a huge difference between companies when it comes to providing meaningful and timely feedback.

Overall, the companies provided absolutely no feedback to 42% of reports, reactions within the required 24 hours came to not even half of reports (46%). Again, only Facebook provided timely feedback to 70% of reports while YouTube remained silent to 97% of reports.

The companies provided feedback to 59% of general user reports. The feedback was provided quite timely with responses within 24 hours (49%) and 48 hours (7%). This means that the platforms provided feedback within two days in more than half of the 432 cases. But this should be at least above 90% based on the rules of the Code of Conduct. Yet, it cannot be ignored that in 41% of cases the companies provided absolutely no feedback to the user reports of cyber hate sent by the partners.

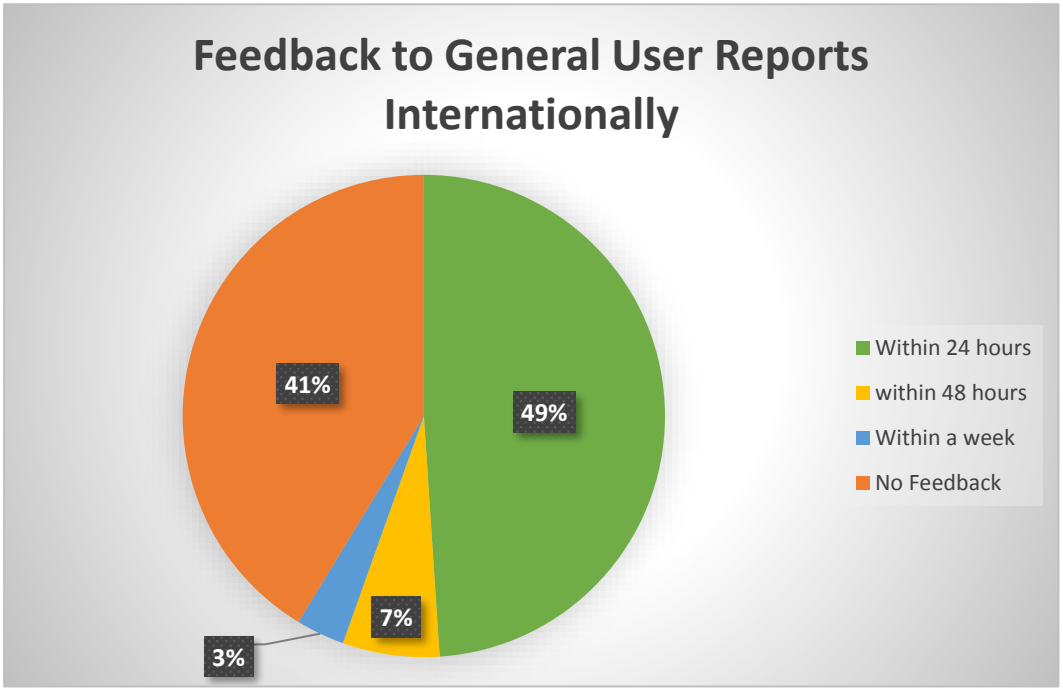

Figure 7: Feedback Times to General User Reports Internationally;
Source: Joint Monitoring INACH/sCAN project

As it has been stated above, companies differ profoundly in this aspect too. Facebook is by far the best in providing feedback. The company responded to the reports within 24 hours in almost 76% of the cases and within 48 hours in 12% of the cases. This means that the company almost reaches the 90% margin in providing substantial feedback in a timely manner.

All the other three companies that the partners monitored during this exercise, even Instagram that is owned by Facebook, are far below Facebook's numbers. Twitter responded within 24 hours in 45% of cases and Instagram responded in 34% of cases within an acceptable timeframe. However, Instagram provided no feedback in 66% of reported cases and YouTube did not provide feedback to any reports sent through general user channels.

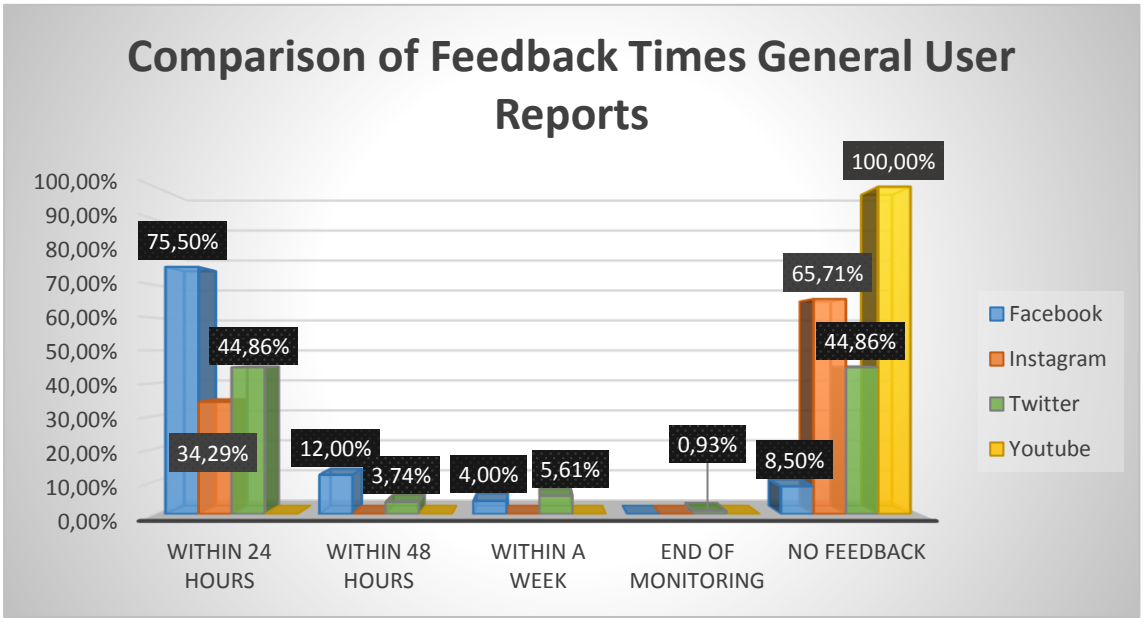

Figure 8: Comparison Feedback Times to general user reports by platform; Source: Joint Monitoring INACH/sCAN project

The situation is not much better when it comes to flagging content via trusted flagger channels. These channels are usually direct email addresses through which our partners can reach members of the platforms' corps of moderators that are higher up the chain. Yet, our partners received absolutely no feedback in almost 50% of cases when reporting through these channels.

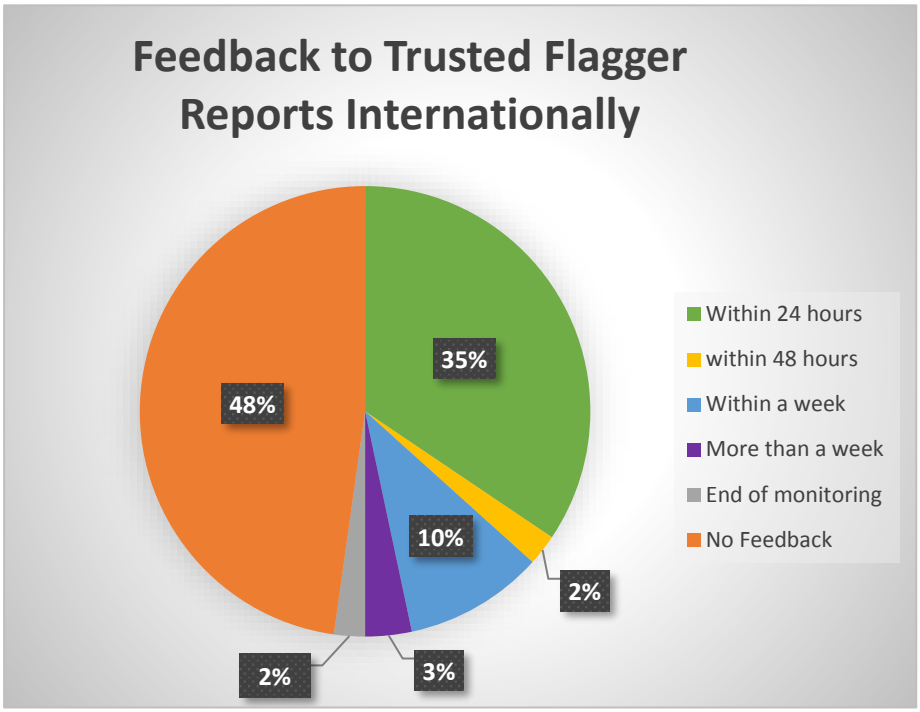

Figure 9: Trusted Feedback Times Internationally; Source: Joint Monitoring INACH/sCAN project

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. Since Facebook received the most reports from the participating organisations, the overall feedback rate for trusted flaggers is lower than that for general user reports.

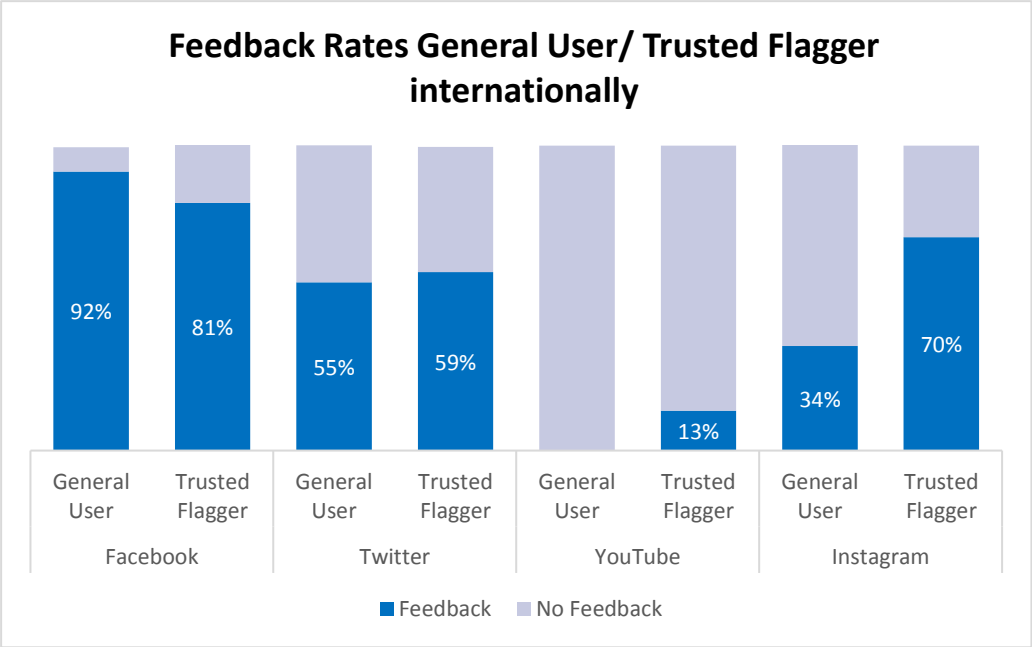

Figure 10: Feedback Rates General User/ Trusted Flagger internationally; Source: Joint Monitoring INACH/sCAN project

The partners experiences during the monitoring

In order to get a better insight into the partner organisations' experiences during the monitoring, an evaluation questionnaire was disseminated at the end 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.

In general, the partners observed that most IT companies reacted less frequently to their notifications than during the previous monitoring exercise organised by the European Commission. In order to effectively combat illegal hate speech online, it is important that social media platforms react to reports from their user base regardless of who is reporting or the time of reporting.

Discussion and Outlook

The International Network Against Cyber Hate (INACH) and the sCAN Project carried out a monitoring exercise – the first one without the oversight and coordination of the European Commission (EC) – to check the adherence of four social media platforms (Facebook, Twitter, YouTube and Instagram) to the Code of Conduct they had signed with the EC. The INACH and sCAN partners reported 432 cases publicly available reporting channels. 90 cases were escalated to trusted flagger channels after having been assessed and rejected by the platforms.

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 general users and trusted flaggers.

Providing no feedback, late feedback or meaningless feedback is a major issue that needs to be addressed by the companies as soon as possible. If people report content that is hateful, discriminatory or inciting violence, it is not enough for platforms to send an automated reply stating that they received the report, or not even that. People should know that these reports are taken seriously so they feel encouraged to report more cyber hate in the future. They should also feel that these platforms are behind them and have their backs if they are being attacked or harassed for who they are.

Moreover, timely and accurate feedback is paramount for NGOs monitoring cyber hate on social media in order to be able to do follow ups and see whether certain content was removed due to their reports or for some other reason.

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 counteraction are pivotal in ensuring a safe and respectful online space across the EU and beyond.