INACH

bringing the online in line with human rights

SCAN

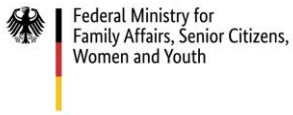
Monitoring Report

20. January – 28. February 2020

International Network Against Cyber Hate - INACH

INACH was founded in 2002 to use intervention and other preventive strategies against cyber hate. 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 its members, the European Commission, the BPB and other donors. The International Network Against Cyber Hate (INACH) currently unites 29 organizations from the EU, 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 educational and preventive strategies.

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



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 20. January 2020 and 28. February 2020, the sCAN project cooperated in organising an unannounced monitoring exercise with the International Network Against Cyber Hate (INACH) and the project Open Code for Hate-Free Communication (OpCode). The goal of the monitoring exercise 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 organisations participating in this monitoring exercise had already been participating in previous monitoring exercises organised by the European Commission and INACH. During this monitoring, the partners reported 484 cases of illegal online hate speech to 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, sCAN partners 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)

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. This monitoring report covers only the cases reported by the sCAN project partners. The results of the monitoring of the OpCode project partners will be published separately.

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

As in the previous monitoring exercises, the methodology followed closely the monitoring process established by the European Commission. 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, or that the companies expressly denied to remove, 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 partners agreed to use a standardised excel template based on suggestions from sCAN partners and prepared by the INACH Secretariat to collect cases in this monitoring exercise.

Key figures

The monitoring took place between 20. January 2020 and 28.February 2020. 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 analysis

To provide an extensive picture of cyber hate within the European Union, the sCAN Project and INACH classified instances of online hate speech during our silent 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. Each collected case of cyber hate could be categorized into as many hate types as it needed to be. So, if a case was, for instance, both homophobic and racist, it was categorised as such. To provide an accurate overall picture, cases that fell into multiple categorise were included in the pie chart below in all those hate types, i.e. they were counted as multiple cases.

[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).

However, this is only true for this section of the report. In other parts of the data, the cases were only counted as one, no matter how many hate types they fell into.

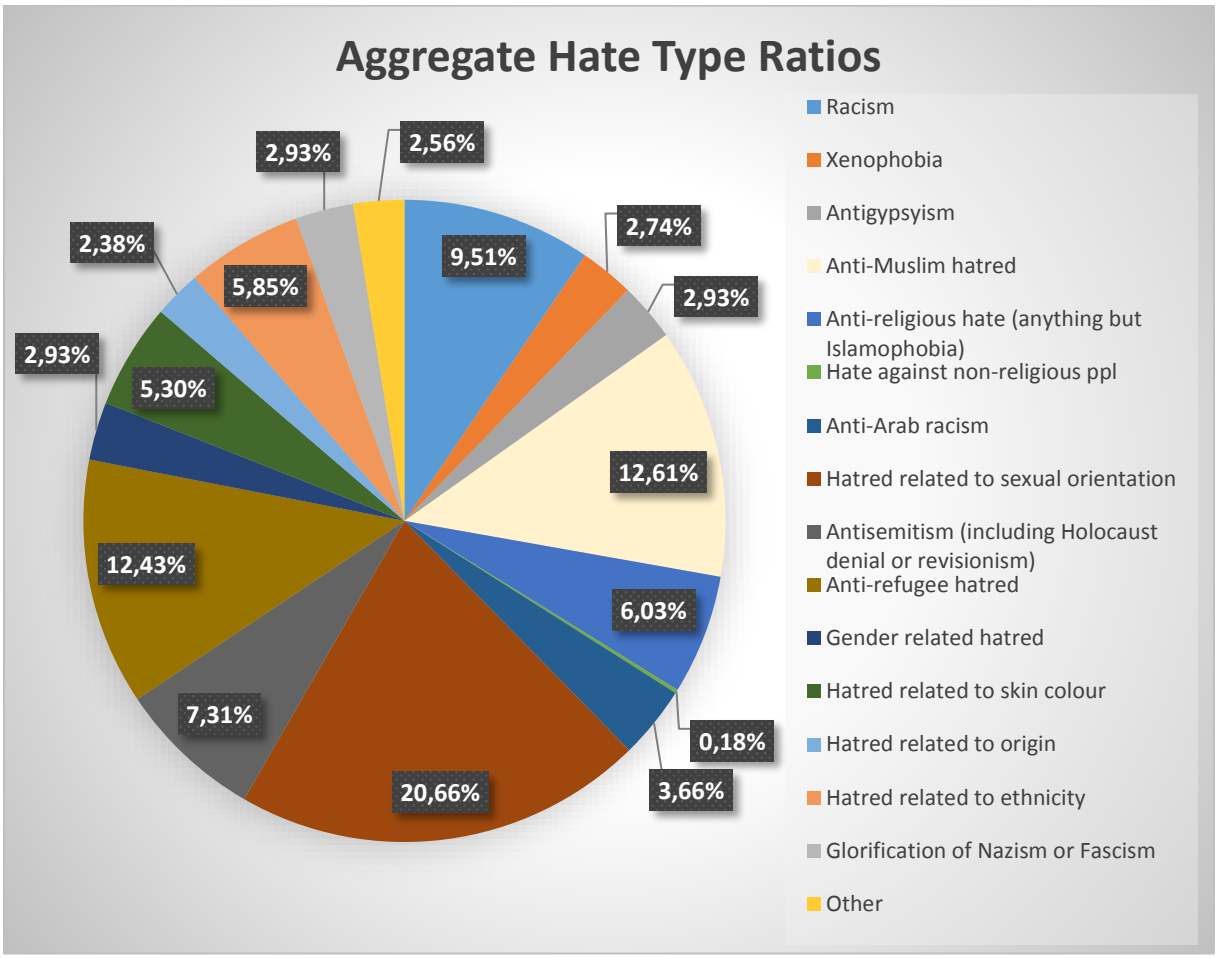


Figure 1: Aggregate hate type ratios

INACH and its members, including almost all sCAN Project partners, have been collecting cyber hate cases for years. Hence, the network has a fairly detailed picture of what hate types are most prevalent in the EU. Naturally, there is always fluctuation in the numbers and there are trends that come and go, but in general: racism, antisemitism and anti-Muslim hatred are the three most prevalent types of hate that we observe. However, this was not the case during this monitoring exercise. Hatred related to sexual orientation (i.e. homophobia) was the most prevalent hate type (20,66 %).

This can be explained by local events in some monitored countries that happened 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. Thus, hatred related to sexual orientation appears to be a major issue in (Central) Eastern Europe in general and in those two countries specifically. 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%). In other words, there is nothing novel or surprising about these findings. These hate types have always been on the top both in INACH's findings and in the Commission's official monitoring exercises. Even hatred related to sexual orientation is usually very prevalent in the findings, if not as staggeringly as in this monitoring due to the aforementioned idiosyncrasies of the sampling.

Removal rates and assessment times

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 results of previous monitoring exercises, including the EC monitoring exercise conducted only a month earlier. These findings highlight the importance of a consistent case handling by the platforms, irrespective of official monitoring exercises organised by the European Commission.

Furthermore, some instances of cyber hate were only removed after having been reported a second time through the partners' trusted flagger channels. In total, 94 cases were escalated through these channels after not being removed by the companies when reported through general user notification channels. To better reflect the difference between cases that were already removed after initial reporting through general user channels and those that were only removed after escalation as trusted flaggers, the figures on removal and feedback will be given separately for the reporting methods.

Reporting as general users

51 % of the cases were 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 14,29 % of cases.

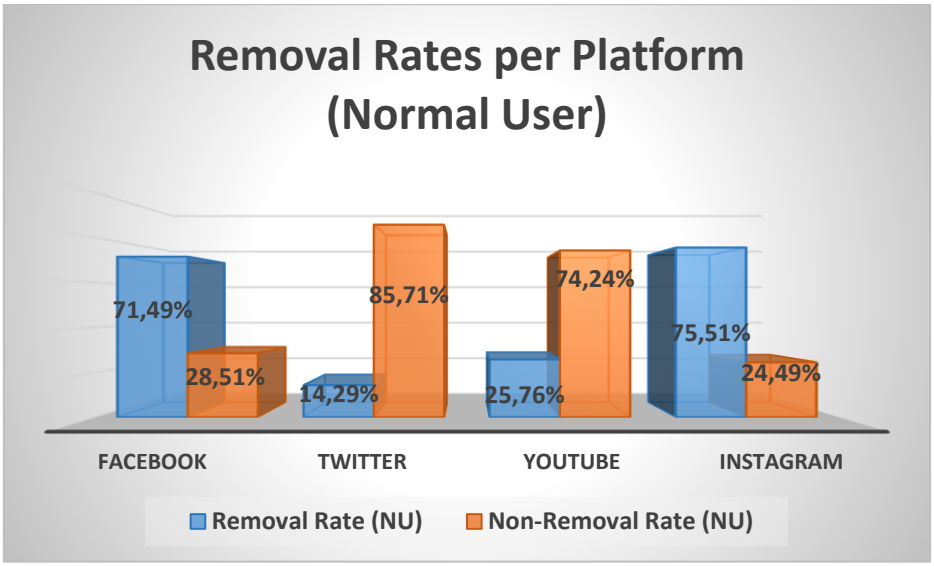


Figure 2: Removal Rates per platform after normal user flagging

In the Code of Conduct, the platforms promise to assess 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 to them 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.

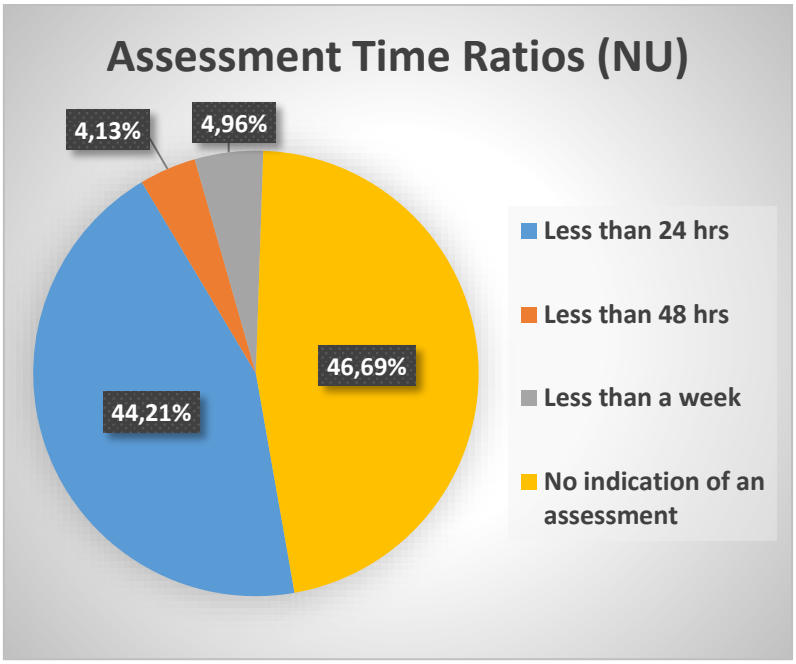


Figure 3: Assessment after normal user notification

Almost three quarters of the cases that were assessed by the IT companies within a week of the notification were removed and the companies 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.

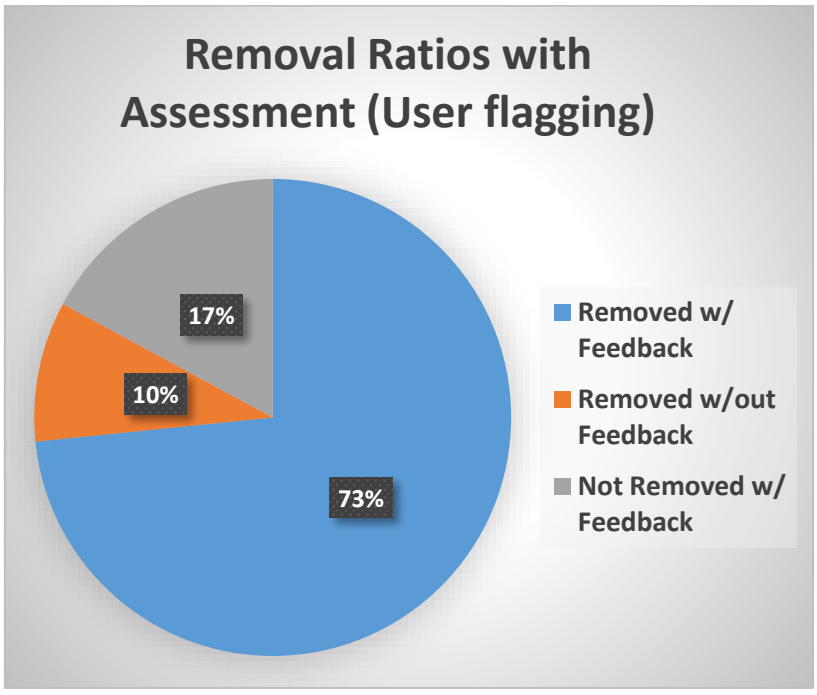


Figure 4: Removal Ratios with assessment after reporting as general users

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 company. 8.18 % of cases were removed at some point between one week after the notification and the end of the monitoring, but the partners were not informed about this removal by the IT companies. It is therefore impossible to tell if the cases were removed as a result of the monitoring or for different reasons.

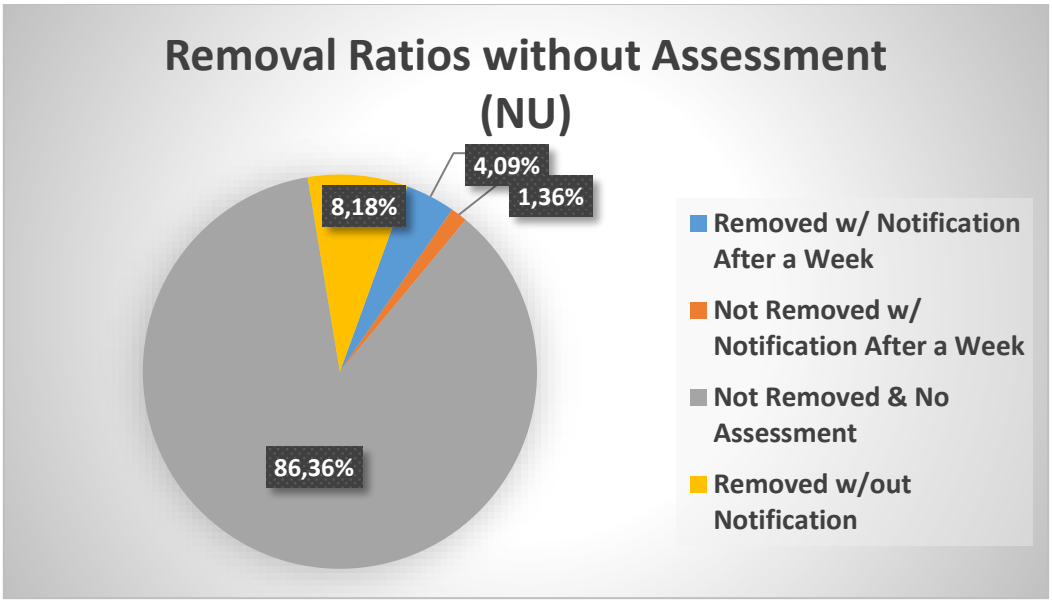


Figure 5: Removal Rations without assessment (normal user flagging)

Escalation through Trusted Flagger channels

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 (46,51 %), while YouTube removed less cases (6,45 %) than when they were reported by general users.

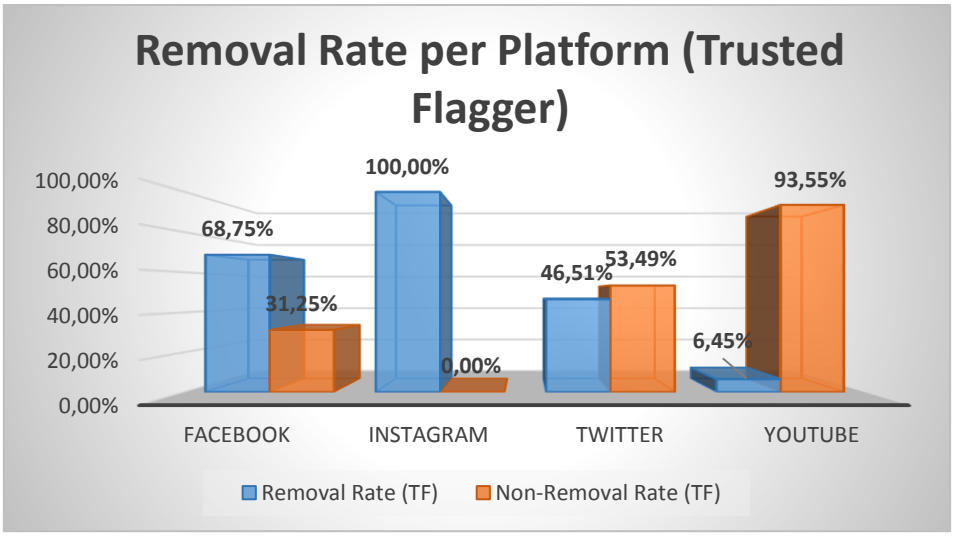


Figure 6: Removal Rates per Platform after trusted flagging

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. These cases were not removed by the IT companies and the reporting organisations did not receive a feedback despite being recognised as trusted flaggers by the companies.

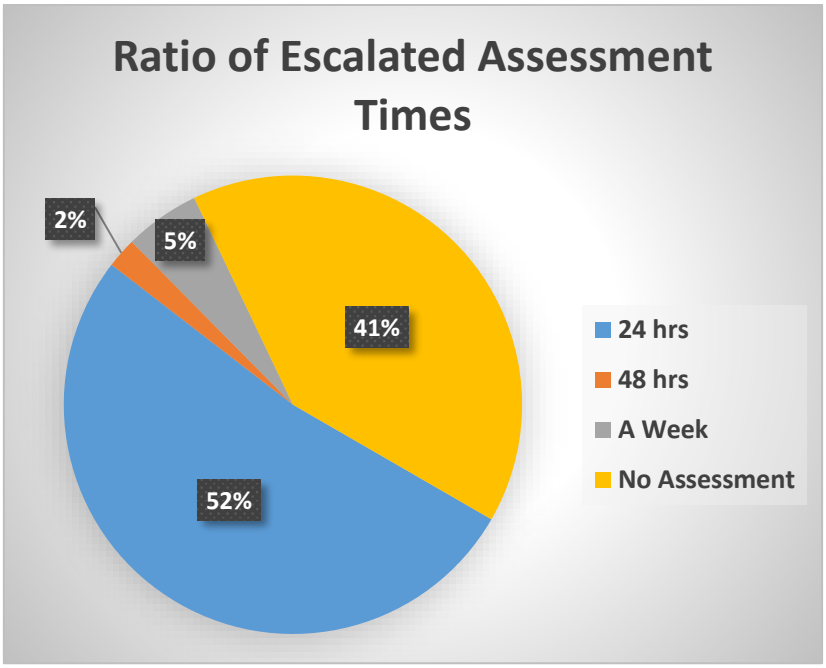


Figure 7: Assessment time of cases escalated through trusted flagger reporting channels

There were huge differences between the platforms 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 being 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 better communication from YouTube about the reported cases in order to enhance cooperation.

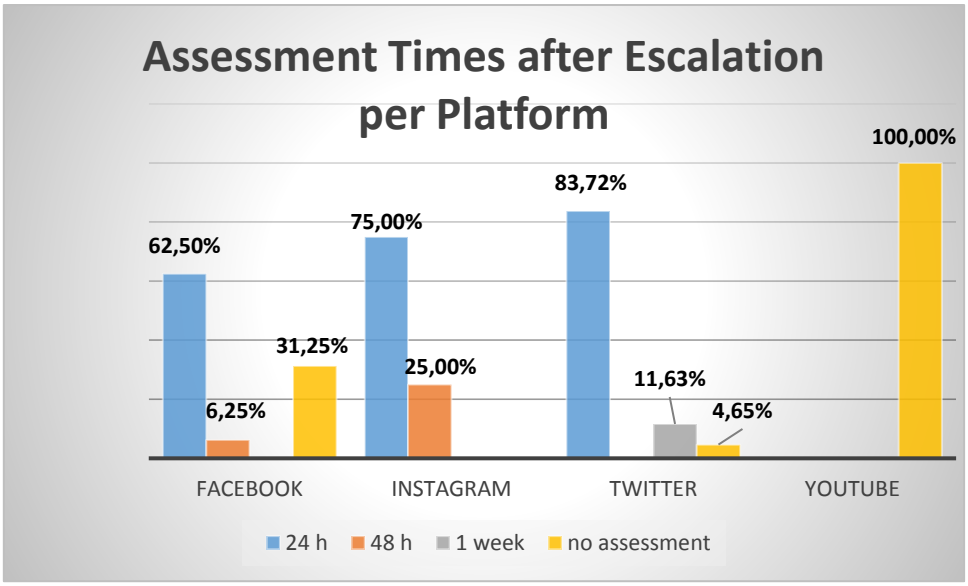


Figure 8: Assessment Times after Escalation per Platform

Country breakdown

One of INACH's main points of criticisms of the monitoring exercises organised by the European Commission is that the Commission does not provide enough detailed data on removal rates and feedback broken down by country. Taking a closer look on the differences between countries is important, as these differences cannot be adequately depicted in aggregated numbers.

In this chapter we will delve into the removal ratios both broken down by country and by platform. The numbers that are examined in this chapter are based on overall removal rates, i.e. the aggregate removal rates of both normal user and trusted flagger cases. A separate chapter will discuss the results of the escalated trusted flagger cases only.

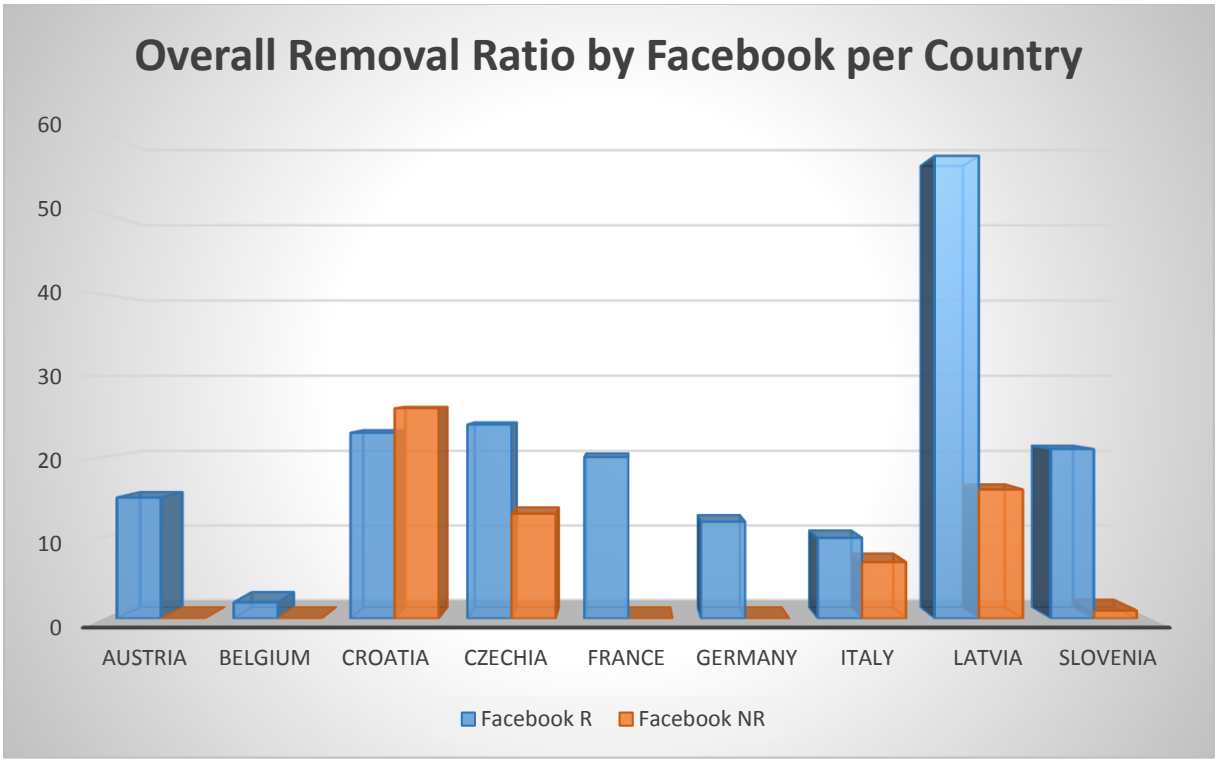


Figure 9: Overall Removal Ratio by Facebook per Country (in number of cases)

It is clear from the data that Facebook had the best overall removal ratio. If we look at the data broken down per country, we can see that this is not surprising. In most participating countries, Facebook's removal rates are quite exceptional. In Austria, Belgium, France and Germany, the company removed all cases reported to them. Even in Latvia and Slovenia, where the platform left some reported cases of cyber hate unremoved, the ratio of removed and not removed cases is very satisfactory. On the other hand, there are still multiple countries where the collected data show Facebook lagging behind. In Croatia, the Czech Republic and Italy, the company did much worse than in the other monitored countries. Thus, the company should focus its efforts more on these EU states to enhance its moderating capacities and reach an EU wide equilibrium in their fight against online hate speech.

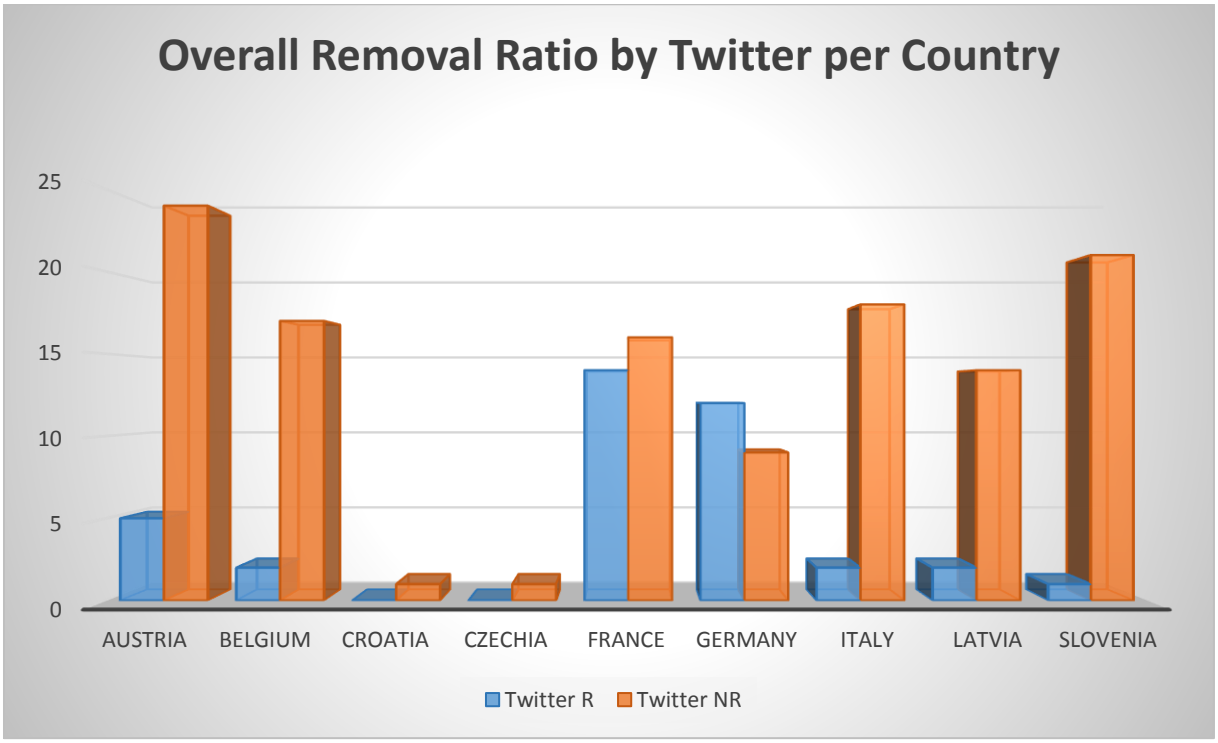


Figure 10: Overall Removal Ratio by Twitter per Country (in number of cases)

Twitter is the only other social media platform that was monitored in all nine participating countries. Compared to Facebook, Twitter performed unsatisfactory in removing instances of cyber hate. There is only one country, Germany, where the company removed more reported cases than not. In France, the platforms removal ratio is just below 50 %. After that there is a major drop, with Twitter only removing about 17 % of reported cases in Austria and it hovers around 10 % in most other monitored countries.

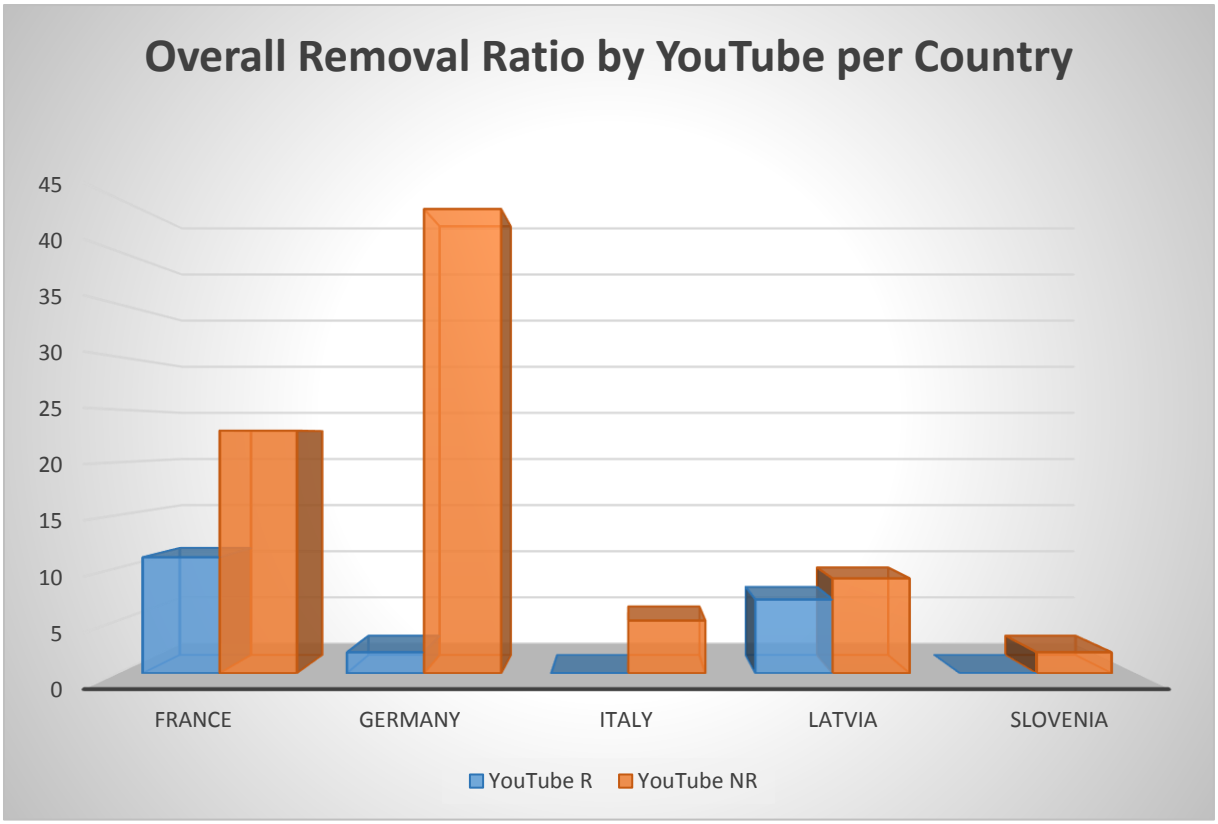


Figure 11: Overall Removal Ratio by YouTube per Country (in number of cases)

Only five of the organisations participating in this monitoring reported cases to YouTube. Just like Twitter, the platform – owned by Google – performed poorly in removing hateful content reported to them. In Italy and Slovenia, the platform did not remove anything. In Germany, they removed only 4% of the reported content. In Latvia, their removal rate was around 43 % and in France around 45 %.

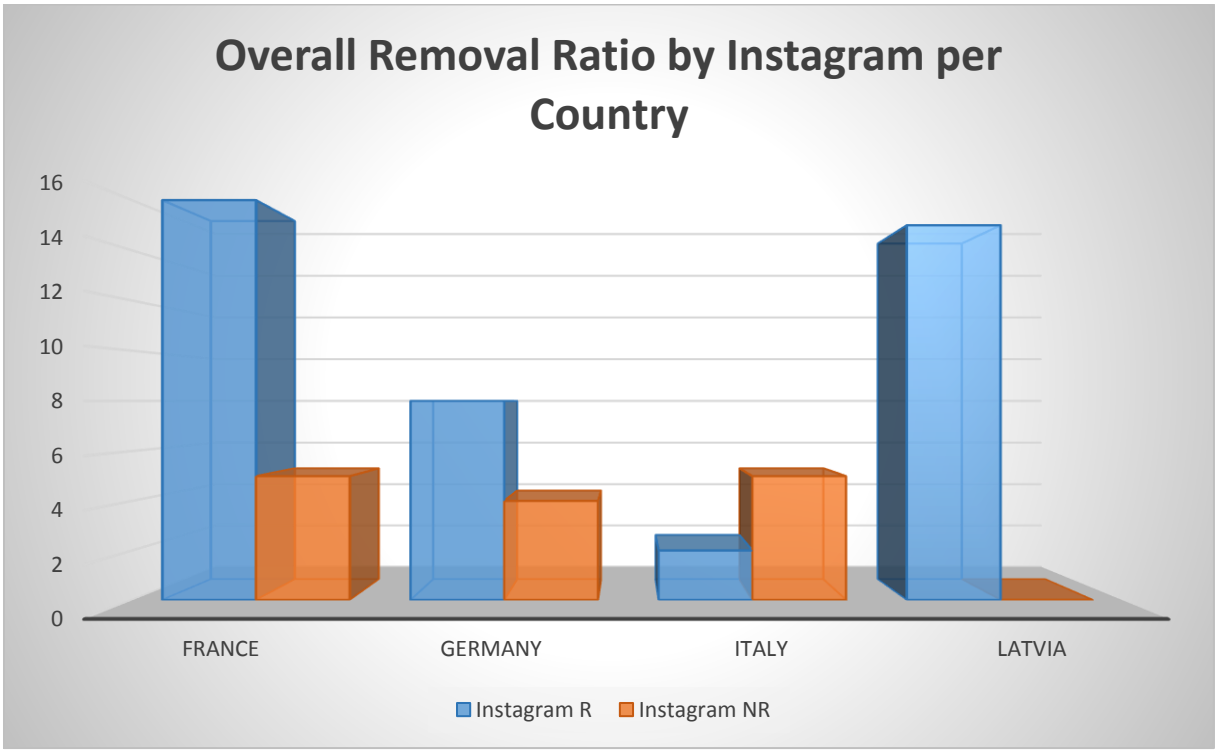


Figure 12: Overall Removal Ratio by Instagram per Country (in number of cases)

Only in four countries did our members report cases to Instagram. Just like its parent company, Facebook, the platform did comparatively well in removing cyber hate. In France, the company removed 76 % of cases, in Germany 66 %, in Latvia, all 15 reported cases were removed by the platform. The only country where Instagram did not remove the majority of reported content was Italy, where only 28 % of cases were removed. This outcome highlights a pattern, confirming the finding that Facebooks moderation efforts are lacking behind in Italy and showing that the company has room for improvement in multiple countries within the EU.

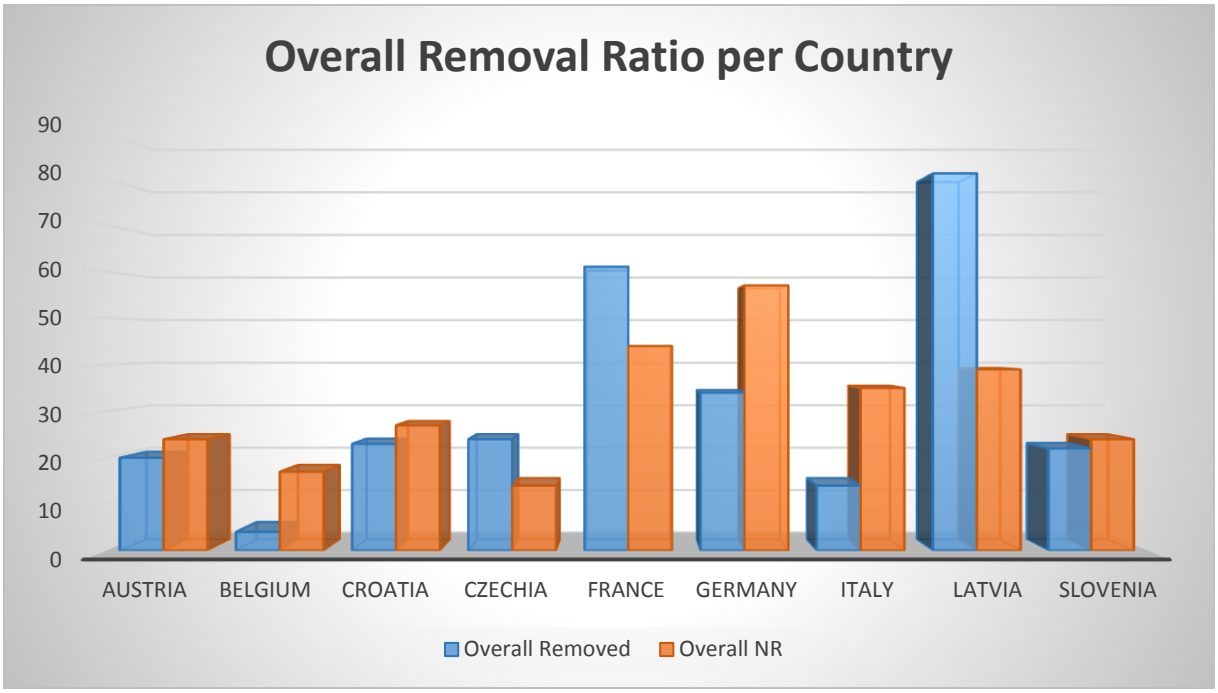


Figure 13: Overall Removal Ratio per Country (in number of cases)

Considering the above described facts, it is not surprising that the overall removal rates in most monitored countries look quite low. There are only three countries, the Czech Republic, France and Latvia, where the companies removed more reported cases than not. Even in these three countries, the removal rate never reaches 70 %. All six other participating countries have very low removal rates, usually hovering around or below 40 %. Without Facebook's particularly good performance on both its platforms, the numbers would show an even grimmer picture. These findings point to the fact that most major social media companies are far from reaching the goals set in the Code of Conduct. Even Facebook and Instagram still have room for improvement in Italy, Croatia and the Czech Republic.

Escalated data by country

Some of our members escalated cases to the companies via so-called trusted flagger channels. These cases were escalated if a case that had been reported via normal user channels was not removed for a week after the notification sent to the platform. The data pictured below is based on these cases, which were not removed after the first complaint made by the monitoring NGOs.

Cases reported via trusted flagger channels are usually considered more closely and they are much more likely to be removed in general. Yet, it is quite clear that the outcomes of these cases during this monitoring exercise are fairly varied.

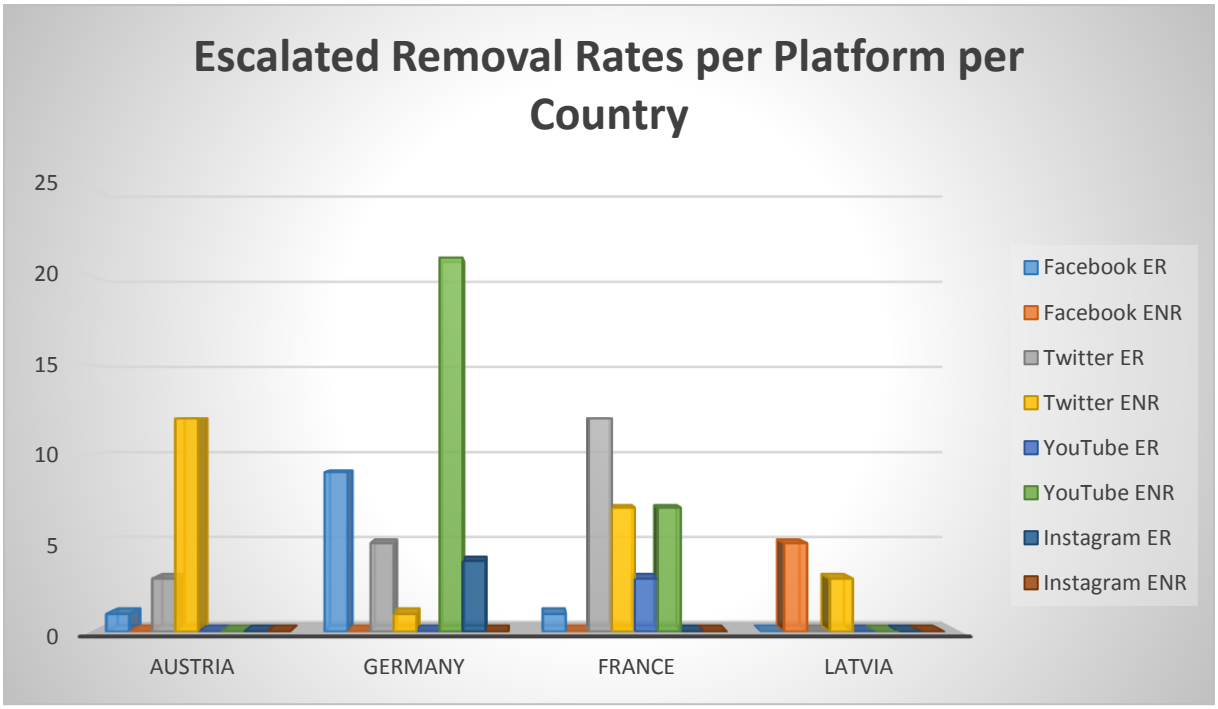


Figure 14: Escalated Removal Rates per Platform per Country (in number of cases)

In Austria, Facebook removed the single escalated case that ZARA sent to them. On the other hand, Twitter only removed 20 % of the cases sent to them by ZARA as a trusted flagger. The picture is fairly similar in Germany, where Facebook removed all cases reported to them by jugendschutz.net as a trusted flagger. Twitter removed 83 % of escalated cases, a much higher removal rate than in Austria. This follows the trend where Twitter did comparatively well in Germany overall. Instagram, just like its parent company, removed all escalated cases. On the other hand, YouTube removed none of the cases reported to them by the German sCAN partner as trusted flagger.

France follows a similar pattern to Germany. Facebook removed everything reported to them by Licra and Twitter's removal of escalated cases is somewhat better (63 %) than their overall removal rate. YouTube removed 30 % of trusted flagger cases in France. Licra did not escalate any cases to Instagram.

The Latvian Centre for Human Rights escalated cases to Facebook and Twitter. Neither of these companies removed any of the cases escalated to them by the trusted flagger. This finding shows – again – that even Facebook still has room for improvement in certain EU countries.

All in all, this separate look at the cases reported by the sCAN Partners to the four companies as trusted flaggers show that there are still discrepancies between normal user and trusted flagger removal. It also shows that there are major inconsistencies between countries even when it comes to trusted flagger reporting. The companies should focus on having a moderation system that works equally for all of their users, regardless where they report from.

Feedback

Receiving feedback on reported hate speech on social media is crucial both for users and trusted flaggers. 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.

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 the findings of 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 10 points of difference for feedback in less than 24 hours. However, in comparison with the last monitoring exercise, the gap has been reduced by half.

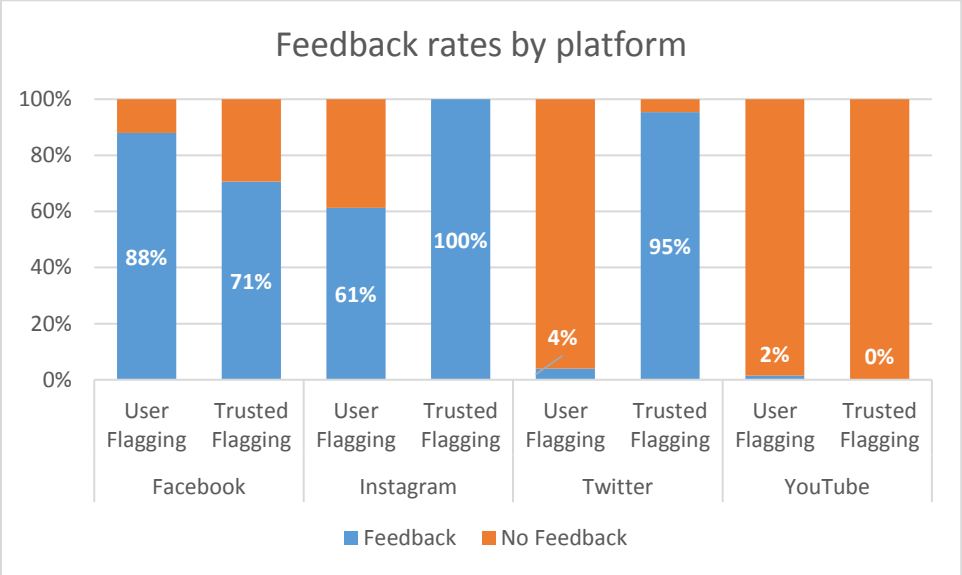


Figure 15: Feedback rates per platform

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 strived to deal better with reporting: user feedback rates almost doubled (from 32 % for the last monitoring exercise to 61 % for this monitoring exercise) and trusted flaggers received feedback in 100 % of the cases.
No significant change was implemented by Twitter or YouTube: feedback rates are still low. Regarding Twitter, the normal user feedback rate has worsened to 11,6 %. The feedback rate for cases reported through trusted flagger channels was 96,6 % during this monitoring exercise.
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 last monitoring exercise. The company did not send feedback to any cases reported by the sCAN partners via trusted flagger channels.
Dealing with reporting took a bit more time for Facebook and Instagram in this 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's feedback time has significantly increased. During the previous monitoring exercise in November, the company provided feedback to normal users within 24 hours in 82% of cases. However, this was only true for 1,57 % cases in this 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 rates are very low.

Conclusion

The results of this monitoring exercise highlight the need for a more consistent performance of IT companies in removing illegal hate speech online across time and location. The overall removal rate of 58 % is almost 10 % lower than the overall removal rate in previous monitoring exercises organised by the European Commission. This includes the 6th EC monitoring exercise in November and December 2019, only one month before this silent monitoring exercise. Companies must ensure that they respond in a timely manner and remove illegal online hate speech at all times. The findings of this monitoring exercise indicate, however, that if a monitoring is run unannounced, the companies achieve much lower removal rates.

The data presented above also show stark differences between the countries monitored in this exercise. While most platforms responded comparatively well in Germany, France and Latvia, the removal rates in the Czech Republic, Croatia and Slovenia were less satisfactory. In Italy, none of the monitored platforms reached the goal of reviewing and removing the majority of illegal content reported to them in less than 24 hours.

Furthermore, the removal rate of the cases reported by the sCAN Partners to the four companies as trusted flaggers show that there are still discrepancies between normal user and trusted flagger removal. It also shows that there are major divergences between countries even when it comes to trusted flagger reporting.

The companies should focus on having a moderation system that works equally for all of their users, no matter where they report from. It is therefore highly important that all monitored companies focus more on enhancing their moderation capacities in (Central) Eastern European Countries and Italy and responding to notifications of their general user base in a timely manner.