INACH

Bringing the Online In Line with Human Rights

Compiled by
Tamas Berecz &
Adinde Schoorl

2023

www.inach.net

Monitoring Report 2023

TABLE OF CONTENTS

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) unites multiple 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.

This publication has been produced with the financial support of the Citizens, Equality, Rights and Values (CERV) Programme of the European Union. The contents of this publication are the sole responsibility of the International Network Against Cyber Hate and can in no way be taken to reflect the views of the European Commission.

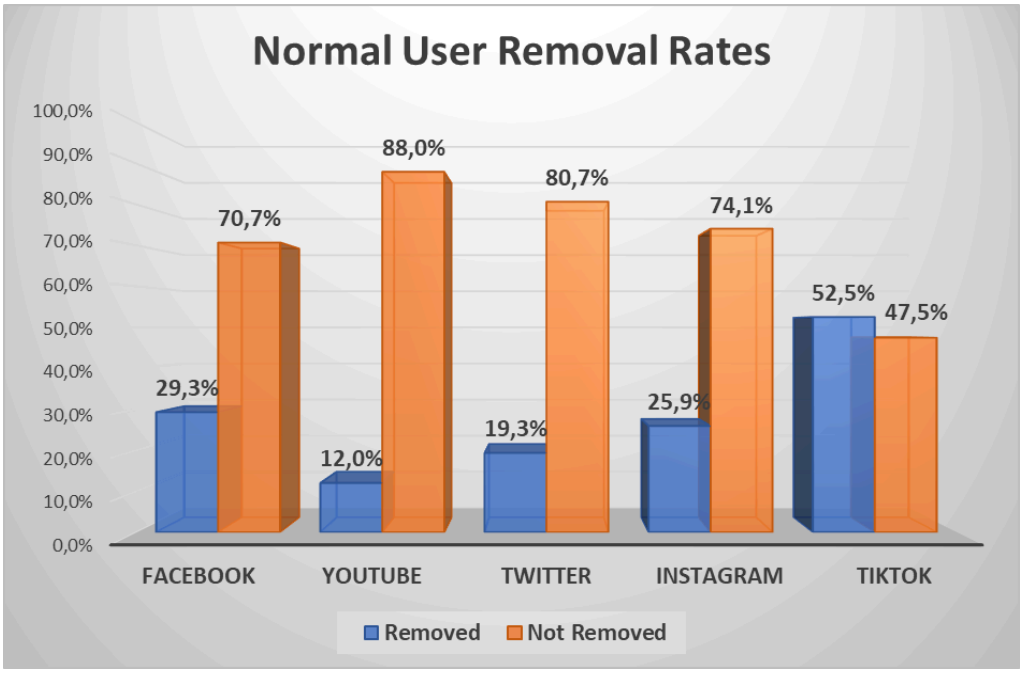
Supported by the Citizens, Equality, Rights and Values (CERV) Programme of the European Union

1.Basic information on the Monitoring Exercise

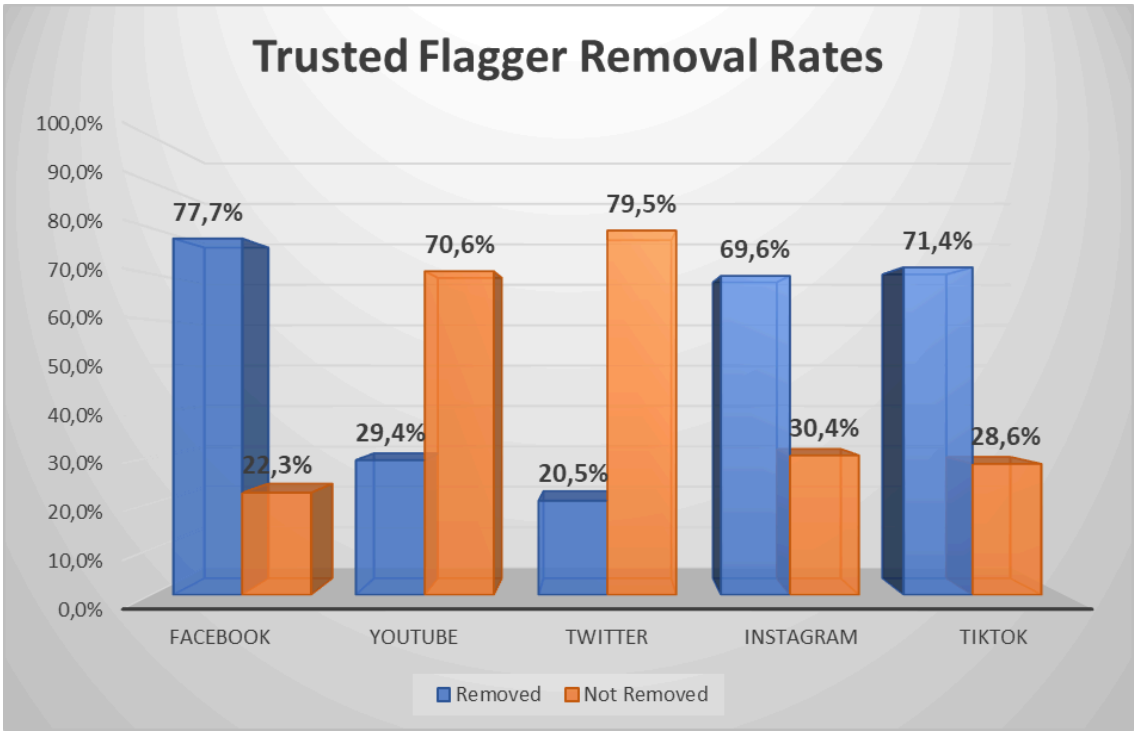
This year the normal annual Monitoring Exercise organized by the European Commission was canceled. However, INACH and other partners organized the Shadow Monitoring Exercise with its partners from the 11th of September until the 20th of October 2023. 20 organizations participated in the Shadow ME from Austria, Bulgaria, Croatia, Czech Republic, Estonia, France, Germany, Greece, Hungary, Italy, Latvia, Lithuania, the Netherlands, Poland, Portugal, Slovakia, Spain and Sweden. More than 2000 cases were gathered during these weeks. The following organizations were part of this ME: CESIE, Human Rights House Zagreb, DigiQ, Dokustelle, Estonian Human Rights Center, FSG, Greek Helsinki Monitor, Háttér Society, ILGA Portugal, INACH, Institute for Law and Internet (Institutet för Juridik och Internet), Integro, Jugendschutz.net, Latvian Center for Human Rights, LGL, LICRA, MCI, Never Again, ROMEA and ZARA.
One final remark before we proceed to the findings of the ME: due to the purpose of our documentation, X (formerly known as Twitter) will still be referred to as Twitter.

2. Findings of the ME

In this section one can find the results of the Normal User Removal Rates, the Trusted Flagger Removal Rates, the Normal User Feedback Rates and Assessment Time Ratios.

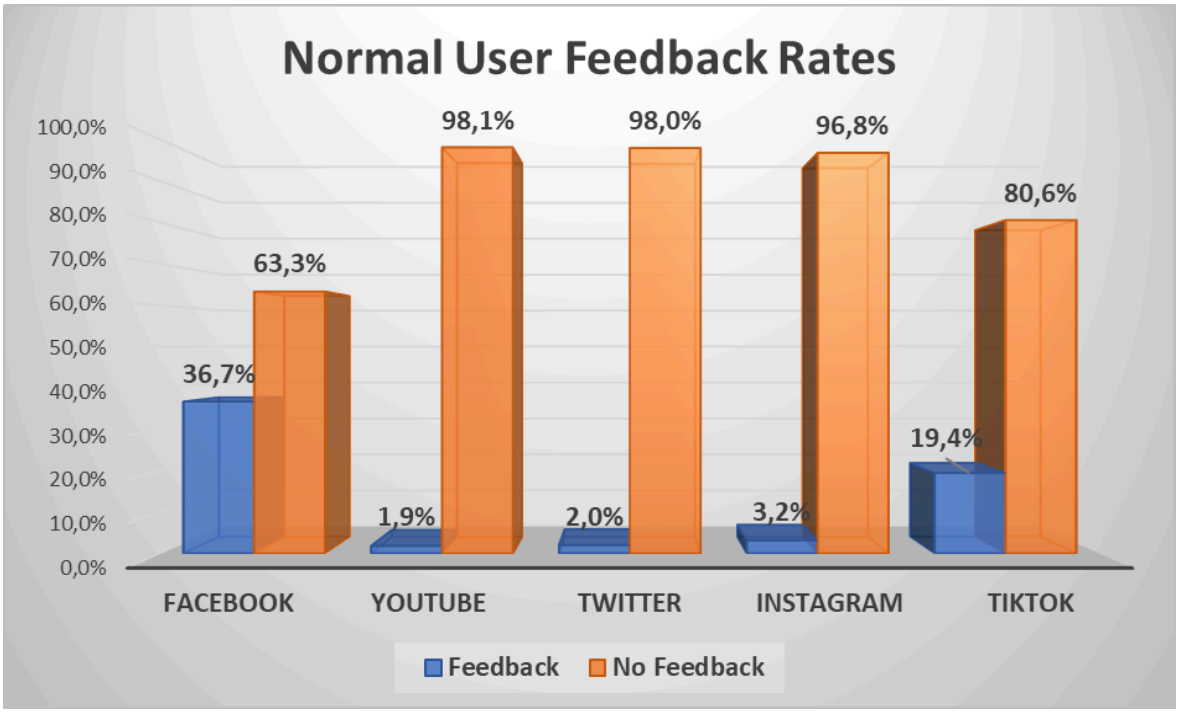


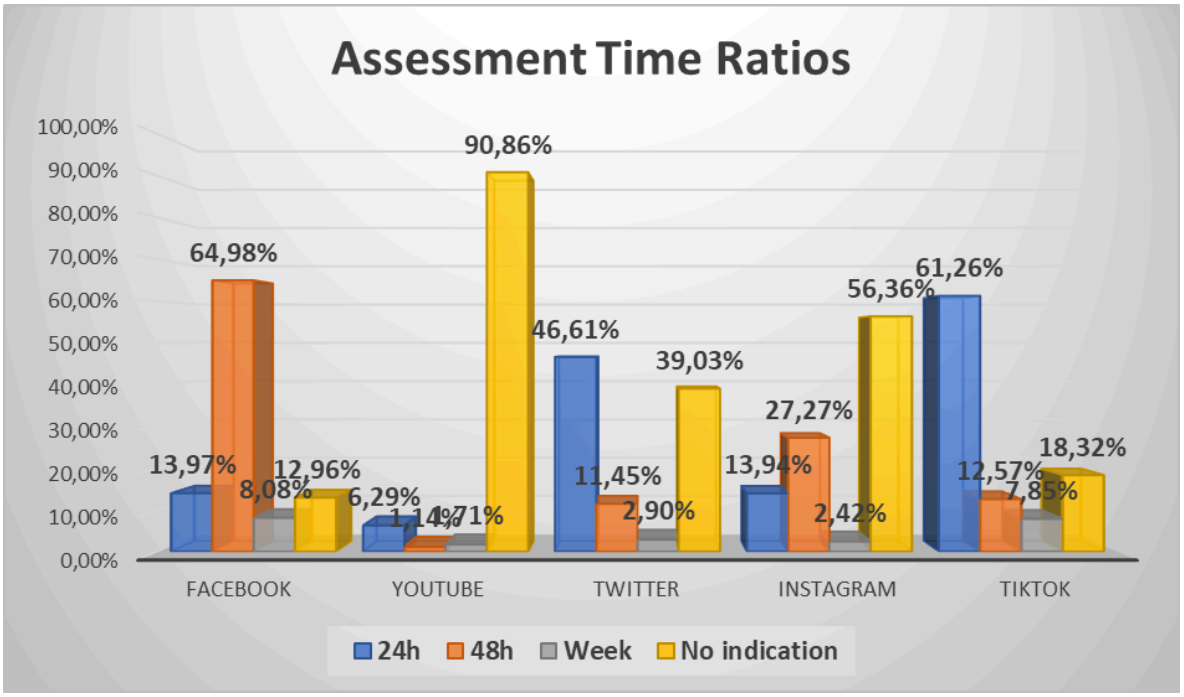
The Normal User Removal Rates show the huge gap between the reported content that was removed and that was not removed. YouTube has the lowest percentage of removing content; only 12% was removed. Twitter removed 19.3%. Instagram and Facebook have similar results between each other: Facebook removed 29% while Instagram has removed 25.9% of the reported content. TikTok turns out to be the only outlier where the removal rate was higher than the 'not removed rate'; 52.5% of the reported content was removed and 47.5% was not removed.



The Trusted Flagger Removal Rate shows actually opposite results; here the removal rates are much higher than the non Removal Rate. Facebook removed 77,7% of the reported content by the Trusted Flagger channel, Instagram removed 69.6% of the reported content and TikTok removed 71.4% of the reported content by Trusted Flaggers. The only outliers are Twitter and YouTube. Through the Trusted Flagger channels they still have a lower removal rate than the 'non removal rate'; YouTube had a removal rate of 29.4% and Twitter had a removal rate of 20.5%.

Third, one can find below the Feedback Ratio for Normal Users. One can straight away see that for the platforms Instagram, Twitter and YouTube the Feedback Ratio is extremely low; 3,2%, 2% and 1.9% respectively. Facebook has the highest Feedback Ratio, with 36.7%. Finally, TikTok has a Feedback Ratio of 19.4%.

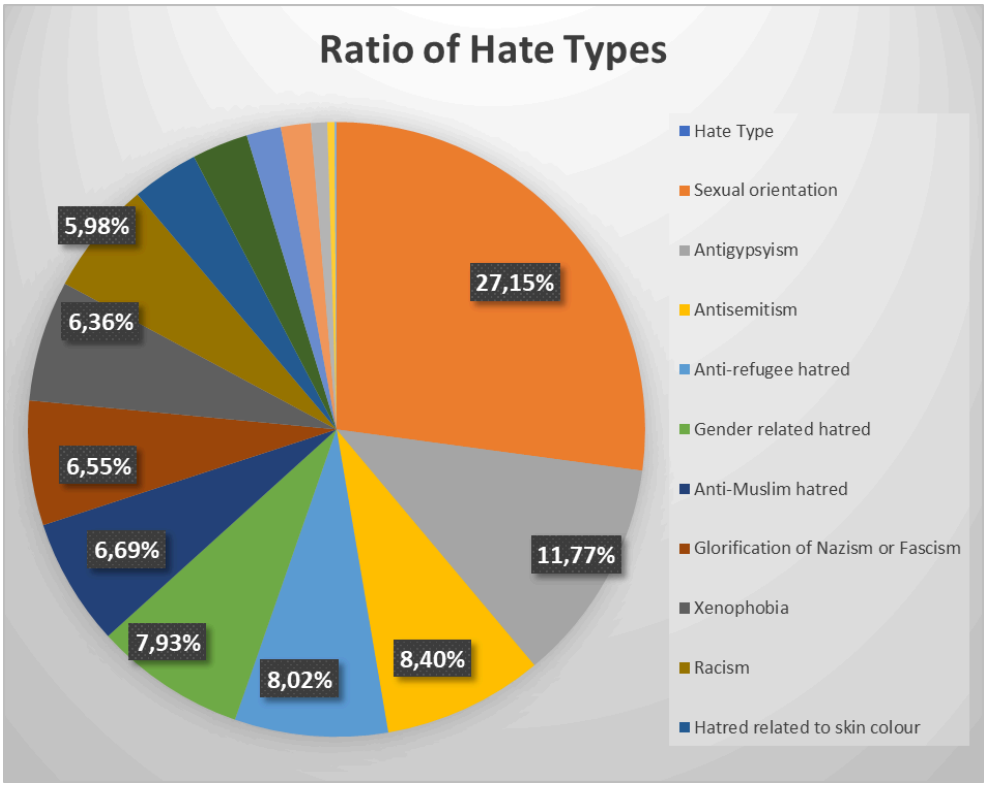




Finally, please find above the results of the Assessment Time Ratios. The results here differ greatly between the platforms. For instance, on YouTube 90.86% of the reported content has no indication of it ever being assessed. On Facebook 64.98% received

assessment within 48 hours while 13,97% received an assessment within 24 hours and only 8.08% within a week. Twitter has a 24 hours Assessment Time Ratio of 46.61%, 11.45% within 48 hours, 2.90% within a week but 39.03% has no indication of any assessment. On Instagram 13.94% received an assessment within 24 hours, 27,27% within 48 hours, 2,42% within a week but 56,36% never had any indication of an assessment. TikTok had the highest 24 hour Time Assessment Ratio: 61,26%. 12,57% received an assessment within 24 hours, 7,85% within a week and only 18,32% never had an indication of an assessment.

3. Types of hate speech and intersectionality



The most prevalent types of hate speech during this Shadow ME were: hatred related to sexual orientation, anti gypsyism and antisemitism. One can see in the graph below

that the percentages of other types of hate speech were quite evenly divided; between 6-8% of the reported content. The hate speech that the organizations focus on, depends on the goals of the organization so it has skewed the data slightly.

Intersectionality often appears when it comes to anti-muslim hate overlapping with anti-refugee hate and racism. A large number of the participating organizations noticed this form of intersectionality. Anti-refugee hate speech exists in large numbers, often mixed with anti-arab hate and anti-muslim hate. Also, in some cases there were intersectionality between anti-LGBT+ hate and racism or anti-LGBT+ hate and antisemitism. There were also cases of intersectionality between Roma people, women and refugees.

4. IT platforms performances and NGO observations

From our qualitative analysis we noticed a few trends. The first one is that according to all 20 participating organizations, the removal rate ranges from extremely low to disappointingly low. The second trend is that the most prevalent types of hate speech were: hatred due to sexual orientation and anti-refugee hatred / racism / hatred related to ethnicity. Other types of hatred were also prevalent e.g. antigypsyism, antisemitism and anti-Muslim hate.
The next trend is the randomness of the communication by platforms; sometimes platforms respond fast, sometimes they respond slow. The decision to remove content also seems random and is probably done by AI. Some content is removed but other - much worse - content is not removed. Sometimes it is communicated that it is removed while it is not, or that the content is restricted in one country without any explanation on why this particular case is excluded from being removed everywhere.

The next trend is that according to the participating NGOs the feedback time was either very poor or differed per platform. From the low amount of removed content most organizations received the feedback within 48 hours or no feedback at all for half of those cases. However, it really also differed per platform; from feedback within 24 hours to no feedback at all. YouTube in general did not ever communicate to any organization. Also, organizations noticed that they only receive automated and standardized responses, no real feedback.
The occurring hate speech is often influenced by current events; the conflict in Israel and the Palestinian Territories, the anniversary of the Bratislava shooting, the recent adoption of equal marriage in Estonia, the earthquake in Morocco or the Quran burning in Sweden. They are also influenced by national political situations such as for example the right wing government in Italy, the openly gay president of Latvia or the elections in Poland. Intersectionality of the online hate content was found when it focuses on connecting hate speech against religion - anti muslim hate -, ethnicity - focusing on people from the MENA region (Middle East and North Africa) - and migrant status, meaning anti refugee hate.

Compared to last year's ME observations, we noticed that the results are more general and there are less differences per country. For instance, last year it was reported that there were quite some regional differences in feedback rates. This year the feedback rate was reported to be very low in all participating countries. Interesting is the observation that last year YouTube was one of the best performing platforms with a very high removal rate. This year participating organizations noticed the lack of communication and impossibility of knowing whether content was removed or not. YouTube performed very low in all analyzing factors.

Platforms can restrict content only in the country where it is reported; 'geoblocking'. Some of the content reported by participating organizations was geoblocked, mostly on X. However, it is impossible to see a pattern that shows why certain content is

geoblocked and why other content is not. It seems to be a random decision by the person who administers the content. The geoblocked content was not different or 'less' harmful than the other reported content that was removed. Sometimes the communication regarding the reported content was not even reliable; content on X was supposed to be restricted for Greece but was in fact restricted for Germany. Other content was supposed to be restricted but in fact was still online.