GA #: 101084457

D-2.4 Annual Report

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.

GA #: 101084457

Project	**SafeNet**
Funding program	**CERV-2022-EQUAL**
Work-package	**2**
Deliverable	**D2.4**
Type:	**R**
Language:	**English**
Date:	**July, 05 2024**
Lead beneficiary:	**LICRA**
Authored by:	**Audrey Koulidiati, LICRA**

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.

GA #: 101084457

Description of the SafeNet project

The 24-month project Monitoring and Reporting for Safer Online Environments seeks to apply a comprehensive and intersectional approach in prevention and fight against intolerance, racism, and xenophobia online. It joins 21 partners, members of the International network against cyber hate (INACH) and the roof organisation itself. Many are trusted flaggers and have taken part in the monitoring exercises within the scope of the Code of Conduct on countering illegal hate speech online. The project will focus on two priorities being 1. continuous monitoring and reporting hate speech content to the IT companies and responsible authorities and 2. awareness raising by regular advocacy towards the social media companies, providing consolidated and interpreted data to national authorities as well as running national bi-monthly information campaigns involving different stakeholders, including IT Companies, public authorities, civil society organisations and media. The project tasks will be organised in 3 work packages consisting of management and organisational framework; monitoring of content deemed illegal under national laws transposing the EU Framework Decision 2008/913/JHA using the methodology from the past monitoring exercises conducted by the European Commission; and dissemination of gathered data to the relevant stakeholders and the general public. Up to 20 000 of cases will be reported, 10 infosheets in English and 170 in other EU languages produced, online training run for the monitoring partners, standards for trusted flaggers reached for all partners, advocacy roundtables and closing conference will be organised. The project fights for targets of online hate based on grounds of racial or ethnic origin, colour, religion, sexual orientation, or gender identity. The second primary target group involves IT companies, national and European authorities, CSOs and media. A wide public will benefit from a kinder internet due to a better and faster removal of hate speech. Project funded by the European Union's CERV-2022-EQUAL.

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.

GA #: 101084457

Table of contents

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.

Lessons from the kick-off meeting

On January 19, 2023, in a 4-hour meeting, the SafeNet project was launched by the INACH network in partnership with 21 EU NGOs.

The majority of SafeNet's consortium already has expertise in reporting illegal online hate speech, corresponding to each partner's local laws. With this project, the consortium - except for the German partner Jugendschutz.net - wants to go further and will also report harmful hate speech. Indeed, the 21 NGOs, members of the SafeNet project, agreed, through votes, on the types of hate speech they were going to deal with during the 2-years project. Almost 90% of participants voted for harmful hate speech to also be treated in addition to illegal hate speech. Also, 66,67% of participants voted to exclude from monitoring the platforms that have not signed the Code of Conduct on Countering Online Hate Speech.
The final project objective is to analyse how IT platforms moderate not only illegal hate speech. To pursue that goal, the consortium will follow the basic guidelines of the EU Framework Decision 2008/913/JHA.

The SafeNet partners also organised social media campaigns as well as the creation and maintenance of the project website in order to disseminate its results. The website is part of INACH's website, ensuring that the project's deliverables remain available after the project has ended.

Training activities

To prepare all participants, INACH organized two 2-hour training sessions on January 25, 2023. These two webinars aimed to present the SafeNet project and answer all the questions the partners might have about its methodology. The coordinators of the project initiated the webinar by presenting the purpose of the project and reminded INACH's official definition of "*hate speech*", "*racism*", "*antisemitism*", "*anti-Muslim racism*", "*antigypsyism,*" and other concepts.

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.

They presented the methodology: how to label and classify the cases. They introduced the data collection and explained how to fill out the data in the INACH platform. They also explained how to be in compliance with the GDPR by removing or blackening all private data except if it is useful to understand the context of the alleged hate speech.

At the end of the webinars, the participants could ask questions and they discussed how to improve this monitoring. Following these training activities, LICRA also provided the definition of the Criteria, Parameters and Definitions Guide available in Appendix 1.

1) Definition of Hate Speech

According to the Committee of Ministers of the Council of Europe, hate speech is understood as « *all types of expression that incite, promote, spread or justify violence, hatred or discrimination against a person or group of persons, or that denigrates them, by reason of their real or attributed personal characteristics or status such as "race", colour, language, religion, nationality, national or ethnic origin, age, disability, sex, gender identity and sexual orientation* ».

Members of the SafeNet consortium also define hate speech as "*intentional or unintentional public discriminatory and/or defamatory statements; intentional incitement to hatred and/or violence and/or segregation based on a person's or a group's real or perceived race, ethnicity, language, nationality, skin colour, religious beliefs or lack thereof, gender, gender identity, sex, sexual orientation, political beliefs, social status, property, birth, age, mental health, disability, disease. Hate speech also includes intentional or unintentional public discriminatory and/or defamatory statements; intentional incitement to hatred and/or violence and/or segregation of a person or persons based on their real or perceived belonging to a certain group or community, or lack thereof. Hate speech also includes intentional and public apologist arguments, negationism, revisionism*

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.

and denial of genocides, especially of the holocaust and other crimes against humanity; or crimes committed by certain oppressive political regimes".

Members of SafeNet's consortium agreed on monitoring online hate speech with an intersectional approach and all types and forms of online hate speech found on IT platforms, including visuals, images, memes and symbols.

2) Reporting of different types of hate speech

The consortium reported the following types of hate speech:

- Racism,
- Anti-Black racism,
- Antisemitism,
- Anti-Muslim hatred,
- Antigypsyism,
- Xenophobia,
- Anti-religious hate,
- Holocaust denial and Holocaust distortion,
- Denial of crimes against humanity,
- Anti-Arab hatred,
- Anti-Asian hatred,
- Anti-refugee hatred,
- Hatred related to skin colour,
- Hatred related to the origins,
- Hatred related to ethnicity,
- Hatred related to sexual orientation,
- Gender-based hatred, including identity and expression,
- Ageism, ableism, and social status hatred,

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.

- Glorification of National Socialism.

However, it is important to note that partners are not obligated to report all the listed types of hate speech above since some specialize in hatred related to sexual orientation or only racist and antisemitic hate speech. SafeNet's partners monitor the intersectional aspect of online hate speech in order to highlight it in their reports; hence it is possible to select different types of hate speech in the database within a single report.

3) List of the platforms for the monitoring

The consortium monitored platforms that have signed the Code of conduct on countering illegal hate speech online[1] which are:

- Facebook,
- Instagram,
- Microsoft,
- Twitter - X,
- TikTok,
- YouTube,
- Snapchat,
- Dailymotion,
- LinkedIn,
- Jeuxvideo.com,
- Rakuten Viber,
- Twitch.

[1] https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.

It is important to note that the partners are not obligated to monitor all platforms listed above. Each monitoring officer prioritises monitoring platforms based on the following criteria: platforms that are relevant in individual countries, have public content (not the content in private or secret groups) and are included in EC Monitoring Exercise.

How to deal with monitoring some categories of hate speech?

Some issues have been raised by partners who shared their concerns about certain categories of hate speech listed above, such as the use of "anti-Muslim racism" instead of "anti-Muslim hatred", or the fact that "Antigypsyism" fails to cover racism against Roma, Sinti, and Travellers.

The issue of the specificity of "anti-Black racism" that can be perceived as not included in the "racism" category has also been highlighted. The project partners agreed to add the two categories when reporting online hate speech targeting black persons.

The main issue shared by many SafeNet members was the social media moderation use of the geolocation and the "visibility limited" issues: On Twitter, most content is withheld rather than removed. The consortium decided to report them as just limited and not removed. This issue has also been reported to the platform during the advocacy activities.

Production of Factsheets and trend analysis

Since the beginning of the SafeNet project, the consortium report 14,899 cases of online hate. In 2023, the consortium report 9,008 cases of online hate. And, in 2024, we can notice that the consortium's partners produce several reports per month:

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.

- January: 936
- February: 1,032
- March: 890
- April: 1,055
- May: 944
- June: 1,185

To analyse these reports, 8 factsheets were produced over the first year of the project implementation - in April, June, August, October, and December 2023 and in February, April and June 2024 - documenting the findings of continuous monitoring of hate speech on social media carried out by the consortium. These factsheets have fewer cases because the data considered dates from June 2024 while the database considers the reports until July 2024.

So, in the factsheets, the consortium reported 13,876 cases of online hate

- Facebook: 5,134,
- Instagram: 1,243,
- Twitter: 5,152,
- YouTube: 800,
- TikTok: 1,547.

4,158 cases have been removed or "withheld on country X", or 32.41%, while 9,718 are still online. 34,38% of all reports received a response within 24 hours from the platforms. There is a significant decline in this rate every month.

During this monitoring period, the following types of hatred were most frequently reported: Sexual orientation: 21 %, Antigypsyism: 11 %, Multiple motives: 11 %, Racism: 11 %, Anti-Muslim hatred: 9 %. The type of hatred most frequently reported depends largely on the country and the specialization of the NGOs. For example, in the Czech Republic, since the beginning of the project, ROMEA has reported 543 hateful posts, the vast majority of which were related to hatred

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.

towards the Roma (66.18%). And in Austria ZARA reported a lot of hate speech against refugees (32,07%).

Moreover, depending on the period, a type of hatred will gain momentum. For example, since October 7, 2023, the rate of antisemitism and anti-Muslim hatred has exploded in Belgium and France.

Dissemination strategy and implementation - social media campaigns

As Coordinator and WP leaders, INACH, LICRA and LGL produced content to allow partners time to familiarize themselves with its concept and implementation. The social media campaign has two elements. First is the presentation of different project partners and each organization's activities. Second, it is the presentation of the local legal context regarding hate speech.

Since the consortium is large and the project envisages using partners' social media channels rather than creating new social media pages for the project, the social media campaign's purpose is to introduce both different partners and different legal backgrounds of hate speech legal framework throughout the EU and present the project with engaging social media posts.

1) The presentation of the project's partners

Every Month, the consortium entered their dissemination efforts into the project dissemination log. It is estimated that the social media campaign has generated 703 posts and reached 182,058 users.

Here are some examples of partners' social media campaign dissemination efforts:

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.

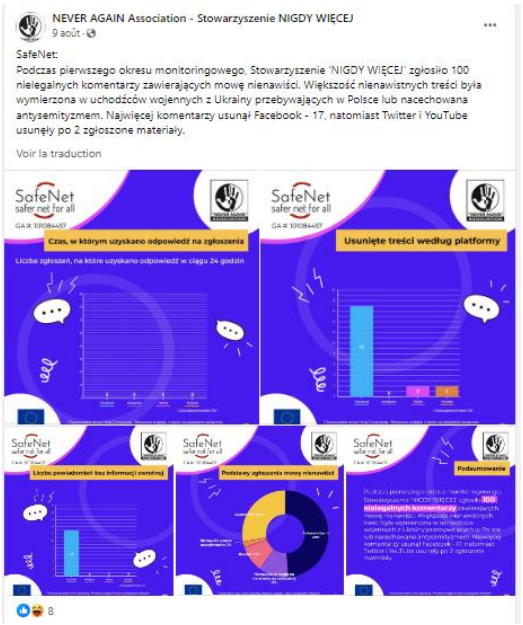
GA #: 101084457

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.

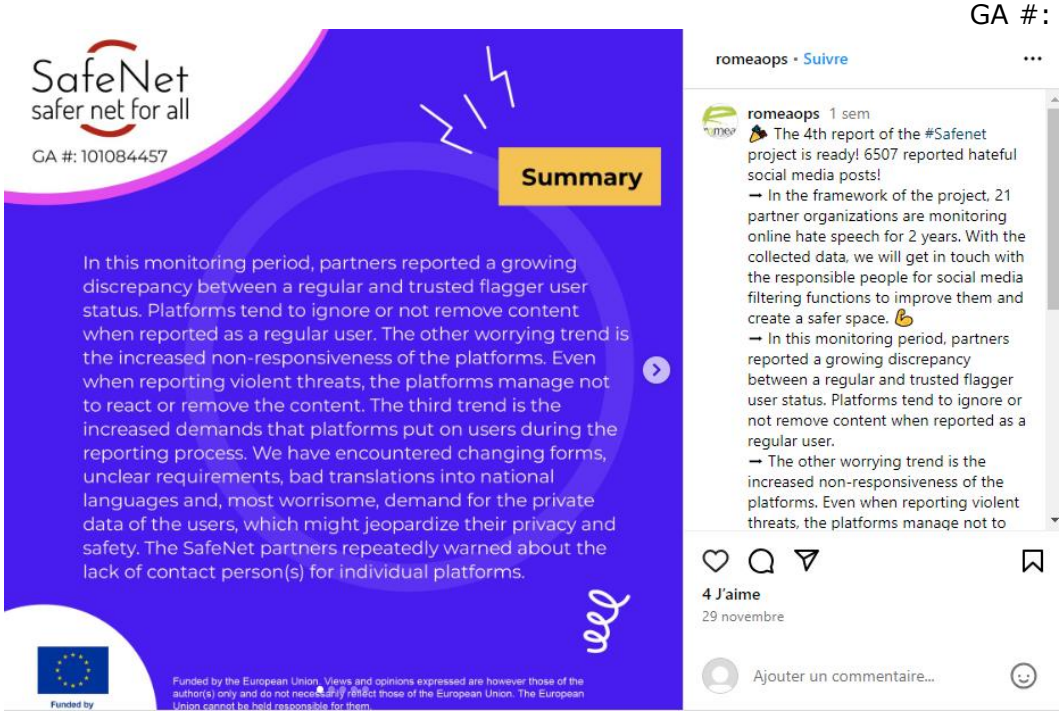
2) The Factsheets dissemination

So far, the dissemination of fact sheets produced 558 posts and generated 151,071 users reached on social media. Here are some examples of fact sheet dissemination on social media:



Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.



3) The creation of a website for the SafeNet project

To make the SafeNet project more accessible and visible, the consortium decided to create a website that is part of INACH's website, ensuring that the deliverables of the project remain available after the project has ended https://www.inach.net/safenet/

Since the site was created on April 1, it has reached 374 users on all continents: some countries in Africa, Asia and South America, as well as all countries in North America and Europe. The SafeNet project website is set to expand, and efforts continue by partners to communicate about it.

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.

Advocacy activities

The advocacy roundtable brought together all the partners and platforms: Microsoft, YouTube, Meta, X (Twitter), TikTok and Viber. The main objective was to discuss their removal policy and their reporting system.

First, the platforms explained their efforts to combat illegal hate speech by updating their community guidelines and outlined how their removal policy of hate speech is managed. Secondly, the consortium presented the monitoring statistics and the issues with geolocation removal.

For some platforms, the consortium expressed dissatisfaction about the fact that there is no feedback about the reports, no way to follow, and no way to see if the content was removed or not. The consortium also noticed that some platforms still allow some Nazi accounts.
Finally, the consortium reminds of the importance of removing hate speech. Indeed, the statistics showed that a great rate of hate content was not removed or was just "withheld in country X" or had a "limited visibility".

A second advocacy roundtable took place on May 29th, 2024. The consortium then noted that there had been no improvement since the first advocacy roundtable. The statistics showed that the rate of removal had not improved and the number of reports with no response from the platforms had increased. The consortium urged the platforms to be more effective in the moderation of online hate speech content.

New challenges with DSA and changing reporting forms

The Digital Services Act aims to regulate the activities of the platforms. The objective is to make digital actors responsible so that they fight against the propagation of illicit, harmful, or illegal content on their services.

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.

DSA will bring new challenges:

To first, enforce in each country all the obligations resulting from the DSA in particular:

- The obligation to denounce the commission of a serious illegal act by a user,
- The implementation of technical and organizational measures to process notifications submitted by trusted reporters is a priority and as quickly as possible,
- The obligation to suspend, after prior warning, for a reasonable period, access to the service to a user who frequently provides manifestly illicit content.

The second challenge is to have effective and efficient sanctions against platforms that do not respect the obligations arising from the DSA. For all these challenges, NGOs and institutions like INACH have their place and the third challenge would therefore be to find how to discuss with the authorities of the Member States and ensure compliance with the DSA.

Conclusion

According to the results of monitoring hate speech on social media, we note that platforms have room for improvement in moderating hateful content online. The percentage of feedback from the platforms is also very low. We are still awaiting the effective transposition of the DSA in the member states, so that we will be able to assess its contribution in the fight against hate content online.

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.