SCAN Platforms, Experts, Tools: Specialised Cyber-Activists Network

Annual Report

May 2018 – June 2019

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

About the Project

The EU-funded project **sCAN** – *Platforms, Experts, Tools: Specialised Cyber-Activists Network* (2018-2020), coordinated by Licra (International League Against Racism and Antisemitism), aims at gathering expertise, tools, methodology and knowledge on cyber hate and developing transnational comprehensive practices for identifying, analysing, reporting and counteracting online hate speech. This project draws on the results of successful European projects already realised, for example the “*Research, Report, Remove project: Countering Cyber-Hate phenomena*” and *“Facing Facts*”, and strives to continue, emphasize and strengthen the initiatives developed by civil society for counteracting hate speech.

Through cross-European cooperation, the project partners will enhance and (further) intensify their fruitful collaboration. The **sCAN** project partners will contribute to selecting and providing relevant automated monitoring tools to improve the detection of hateful content. Another key aspect of **sCAN** will be the strengthening of the monitoring actions (e.g. the monitoring exercises) set up by the European Commission. The project partners will also jointly gather knowledge and findings to better identify, explain and understand trends of cyber hate at a transnational level. Furthermore, this project aims to develop cross-European capacity by providing e-learning courses for cyber-activists, moderators and tutors through the Facing Facts Online platform.

sCAN will be implemented by ten different European partners, namely ZARA – Zivilcourage und Anti-Rassismus-Arbeit from Austria, CEJI-A Jewish contribution to an inclusive Europe from Belgium, Human Rights House Zagreb from Croatia, Romea from Czech Republic, Licra – International League Against Racism and Antisemitism and Respect Zone from France, jugendschutz.net from Germany, CESIE from Italy, Latvian Centre For Human Rights from Latvia and the University of Ljubljana, Faculty of Social Sciences from Slovenia.

The sCAN project is funded by the European Commission Directorate – General for Justice and Consumers, within the framework of the Rights, Equality and Citizenship (REC) Programme of the European Union.

Legal Disclaimer

This Annual Report was funded by the European Union’s Rights, Equality and Citizenship Programme (2014-2020).

The content of the Annual Report represents the views of the authors only and is the sole responsibility of the sCAN project consortium. The European Commission does not accept any responsibility for use that may be made of the information it contains.

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

Content

Introduction

The internet is an integral part of everyday communication worldwide. While it is most often used in a peaceful manner to communicate with friends or freely express ones opinion on a diverse range of topics, some users spread hatred and incite to violence against vulnerable minorities. The international nature of the internet and its global interconnectedness necessitate a transnational approach to combat hate speech online.

In recent years, several European projects countering hate speech have already been successfully implemented. To strengthen European networking and to harness synergies between different projects' results, the sCAN project closely cooperates with the International Network Against Cyber Hate (INACH) and the Facing Facts! project.

The project partners have agreed on INACH's definition of hate speech:

> ***"Hate speech is 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"***[1]

In order to effectively counter-act cyber hate, a multidimensional approach is needed. The sCAN project, therefore, was designed to tackle three areas deemed crucial in this endeavour. This Annual Report presents the sCAN project's results for the first project year (May 2018 – June 2019) in the three areas of combating hate speech covered by the project's activities.

Firstly, to understand the complex issues and phenomena involved in the field of hate speech and counter-action, the partner organisations gathered tools and pooled resources across the countries covered by the project.

Secondly, joint research projects were carried out to achieve an in-depth understanding and analyse transnational trends of different hate speech phenomena. Furthermore, the sCAN partners participated in the fourth monitoring exercise set up by the European Commission to evaluate IT companies' adherence to the Code of Conduct on countering illegal hate speech online. A second monitoring period was carried out by the sCAN partners together with INACH. Supplementary to the thematic research projects and joint monitoring exercises, sCAN partners also shared their observations on hate speech developments in their respective countries to keep up-to-date with trends and draw conclusions of common themes.

Research and monitoring alone are, however, not sufficient to combat cyber hate. Besides reporting illegal content to social media companies and website providers, the partners developed online courses and offline training to build capacities of NGOs and individual activists to counter hate speech in a diverse manner, through monitoring, engaging in counter speech or moderating online discussions.

[1] International Network Against Cyber Hate (2018). *What is cyber hate*. Available at http://www.inach.net/wp-content/uploads/WHAT-IS-CYBER-HATE-update.pdf (last accessed 24.06.2019).

Tools and Resources

To facilitate monitoring and research efforts, the sCAN partners took stock of already existing resources and tools. As many of the tools work with keywords to identify hate speech, the sCAN partners compiled a dataset of keywords in all project languages, including additional information on the context in which those words are used in the respective national discourses.

Additionally, the project produced a study mapping available software solutions to automatically monitor cyber hate. Those tools are important for researchers to keep up with the ever growing amount of (hateful) content disseminated on the internet. Initial testing of the identified tools was carried out and will be consolidated during the project. Results of the testing will be published in the second year of the project.

Hate Ontology

The organisations participating in the sCAN project have extensive knowledge about the particulars of hate speech and hateful online discourse in their respective countries. This includes knowledge of certain words and language often used in those countries to promote discrimination and hatred towards vulnerable minorities. Keywords alone, however, are not sufficient indicators for hate speech. It is equally important to keep in mind the context of a post or comment.

In order to provide an overview of words and codes indicating hate speech and a guidance to the possible contexts they are used in, the sCAN project partners developed a Hate Ontology. This ontology covers the categories racism and xenophobia, antisemitism, anti-Muslim hatred, anti-refugees hatred, antigypsyism, homophobia, misogyny, hate against disabled people and hate against socially disadvantaged groups in the languages of the project countries: French, Czech, Croatian, German, Italian, Slovenian and Latvian. When applicable, references to the historical, cultural and social origins and/or context of use of the reported terms and expressions are provided.

A common trait in racist discourse in the analysed countries is a general fear of so called “racial mixing”, loss of national identity, traditions and values. This is well reflected by nationalist slogans like “Germany to the Germans”/” Slovenia to the Slovenes”/ ”Italy to the Italians”, etc. which are common among several countries.

Intersectional hate speech about migrants, refugees and Muslims was also observed in several countries. There appears to be a widespread equation of refugees and Muslims, leading to all Muslims being perceived as refugees and vice versa. A common trope is the fear of a so-called “Islamisation” of European societies.

Conspiracy theories, holocaust denial and prejudices against Jews are common in all countries and are sometimes even employed unconsciously by people otherwise not demonstrating antisemitic sentiments.

Homophobic hate speech surges in conjunction with national coverage of events like Pride Parades or social and legal improvements for the LGBTIQ community.

Misogyny emerges mostly, when a woman with national visibility publicly advocates and expresses support to a given cause, regardless the topic. These attacks are often targeting the physical aspect and alleged behaviour of the target, even when those are of no relevance for the given debate.

Derogatory expressions about people living with mental or physical disabilities are often used to defame or mock people who do not bear disabilities themselves. This is a stark prove of both the prevalence but also the acceptance of prejudices against disabled people in the general society.

The findings of the Hate Ontology provide valuable insights into the nature of hate speech and hateful discourse in the analysed countries. Apart from providing a starting point to the monitoring and contextual evaluation of hate speech it also serves as a basis for further discussion on the prevalence of discriminatory prejudices in societies at large.

Mapping study

NGOs monitoring internet platforms for hate content have traditionally relied on human researchers manually searching and reporting cases or engaging in counter speech. With the rise of social media, the amount of online content produced every day has become largely uncontrollable by human moderators alone, leading to the development of a number of technological solutions to facilitate this work. Not all of those solutions are, however, easily accessible to civil society for various reasons.

sCAN experts identified a range of software solutions and tools to automatically scan online content for keywords indicating hate speech in the various national contexts. Those tools usually employ technology known as web 'scraping' or 'crawling', automated scripts browsing websites or social media platforms for a set of predefined keywords and structuring the data in a database or spreadsheet. The different crawlers available differ in their functionality and the platforms they can be applied on. A common feature of such solutions is the huge amount of raw data output, which then has to be manually sorted by human researchers who can identify the national contexts.

In recent years, IT companies have developed Artificial Intelligence (AI) technology to facilitate monitoring and content moderation. Importantly, AI is a general term comprising diverse applications like expert systems, machine learning and deep-learning technologies. Expert systems are programmed to make expert-decisions in real-life situations and propose solutions to complex issues. Machine learning is an application of artificial intelligence (AI) based on algorithms that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning tools often involve Artificial Neural Networks (ANN), brain-inspired systems which simulate intelligence by replicating a human learning process. Deep learning algorithms go a step further by attempting to model high-level abstractions in data to determine meaning.

AI is often developed and used by social media companies themselves and tailored to their specific moderation needs. Those solutions are not accessible to civil society. However, start-ups and private companies are also developing AI systems that can be applied in monitoring and moderating hate speech. Human resources are needed in AI development to program and train the algorithms. Especially in the field of hate speech monitoring, human expertise is furthermore needed to assess the algorithms' performance in identifying hate content correctly across various national contexts and languages.

For NGOs, several considerations can influence or hamper the choice of automated software tools to facilitate their monitoring work. Firstly, those tools are usually quite costly and tools which are available free of charge only offer limited functionalities. In order to gain access to all functionalities, users have to subscribe to a paid premium version. AI solutions are more expensive than crawlers, as their development and training requires significant human and financial resources. Most NGOs do not have sufficient funding and resources to employ expensive technology.

Secondly, the tools currently available cannot be applied across all social media networks but have been programmed to work on specific platforms only. The emergence of new social media platforms makes it necessary to keep developing AIs, crawlers or software exclusively targeting contents of the selected platform. NGOs monitoring multiple social media platforms and traditional websites would therefore have to work with different tools for each platform.

Some limitations to the work with automated monitoring tools are also rooted in their functionality. Crawlers need to be supplied with a specific set of keywords taking into account national context and peculiarities of hateful discourse in each country. Apart from the occurrence of those keywords in non-hateful contexts, the constant evolving of slang and slur words necessitates a constant re-evaluation of those sets. Furthermore, it is relatively easy to evade detection of hate speech by tools based on keywords. For example, the use of codes, abbreviations or the altering of text by including numbers or deleting spaces between the words can seriously hamper the effectiveness of those tools. AI algorithms are furthermore subject to several human biases, which can be unintentionally incorporated during their development and training by the choice of training material or the categorisation of the material by human researchers.

Nevertheless, automated technology can be a useful tool to supplement NGOs work on monitoring hate speech. It is highly unlikely to replace human resources in this field completely, but a combination of human researchers and software solutions can reduce the workload and lead to a more effective monitoring. Cooperation between IT companies and NGOs is important to combine technological knowledge on the development of AI algorithms with the expert knowledge on ever evolving trends, context and language in hate speech.

Monitoring and Research

In order to gather in-depth knowledge and enable transnational comparisons on specific hate speech phenomena, the sCAN partners carried out joint research projects on antigypsyism on the internet and alternative platforms for online hate speech – beyond Facebook, Twitter and YouTube.

Furthermore, the sCAN partners participated in the European Commission's fourth monitoring exercise on the Code of Conduct on countering illegal hate speech online and carried out a joint monitoring exercise together with INACH. The objective of the monitoring exercises is to test IT companies' reaction to reports on illegal hate speech on their platforms. The sCAN partners evaluated the actions taken by the tested IT companies Facebook, Twitter, YouTube and Instagram, as well as whether the companies provided feedback about their action to the reporters.

Additionally, the project partners shared their knowledge on current trends regarding hate speech in their respective countries, the tools most commonly used to disseminate hateful messages and online or offline event sparking cyber hate.

Analytical Paper "Antigypsyism on the Internet"

A transnational understanding of the phenomenon of online antigypsyism is necessary to devise effective strategies to counter it. The sCAN partners have chosen the definition of the Alliance against Antigypsyism as a common basis for their research on the subject. The Alliance defines antigypsyism as follows:

> ***"Antigypsyism is a historically constructed, persistent complex of customary racism against social groups identified under the stigma 'gypsy' or other related terms, and in-corporates:***
>
> ***1. a homogenizing and essentializing perception and description of these groups;***
>
> ***2. the attribution of specific characteristics to them;***
>
> ***3. discriminating social structures and violent practices that emerge against that back-ground, which have a degrading and ostracizing effect and which reproduce structural disadvantages."***[2]

Even though several countries have recognised Roma and Sinti as a national minority, the historically continuous hostility, the history of systematic persecution and deeply embedded stereotypes continue to severely impact the lives of people perceived as 'gypsies'. In line with the omnipresence of internet and social media in every day communication around the globe, most antigypsyist rhetoric takes place online. Online antigypsyism is not only spread by far-right and right-wing extremist actors, but widely accepted in the general public and disseminated by political parties, individual politicians and the media.

The main narratives of antigypsyism online mirror the historical stereotypes and narratives that have been used for discrimination and persecution of Romani and other communities perceived as 'gypsies' for centuries. Criminalisation and construction of Sinti and Romani people as 'beggars' serve as excuses to call for discriminatory treatment and exclusion from the social aid system. Interestingly, the notion of 'travelling communities' remains a widespread stereotype, despite the

[2] Alliance against Antigypsyism (2017). Antigypsyism - A Reference Paper. p. 5. Available at http://an-tigypsy-ism.eu/wp-content/uploads/2017/07/Antigypsyism-reference-paper-16.06.2017.pdf (last ac-cessed 24.06.2019).

majority of Sinti and Romani people living a sedentary life. The de-humanisation expressed in many comments on Social Media platforms and online media outlets often leads to calls for violence and even genocide. Fake news and the de-contextualisation of images and videos is a popular tool to disseminate antigypsyist narratives and incite hostility against Sinti and Romani people.

Social Media, especially Facebook, YouTube and Twitter, remain the major distribution channels for antigypsyist hate speech. Discussions in comment sections of YouTube videos and beneath the articles of online media outlets often become platforms for de-humanisation and incitement to violence. Biased media reporting reinforces existing negative stereotypes. A special responsibility also lies with politicians and other public figures.

In order to combat antigypsyism efficiently, Civil Society Organisations need to cooperate more strongly with Romani representatives, Internet Service Providers and public authorities. Media should take care to provide unbiased reporting on Sinti and Roma as well as other marginalised minorities. Reliable moderation is needed in online discussion forums and the comment sections of online media outlets in order to prevent hateful content from reproducing hostilities and dominating the discussions.

Analytical Paper "Beyond the 'Big Three' – Alternative platforms for online hate speech"

Even though the social media giants Facebook, Twitter and YouTube are most often mentioned in research studies on online hate speech, other platforms are gaining importance especially among young users. Instagram, for example, is already more popular than Facebook among internet users under the age of 30. Our research showed that hate groups and extremists wishing to influence minors or young adults with their ideologies follow their target group to those platforms.

Other platforms, like VK.com or Gab.ai are used as alternative platforms or 'safe havens' for hate groups or extremist individuals whose profiles have been suspended on mainstream social media. What makes those platforms appealing to their target group is their more lenient community guidelines and moderation policies towards hate speech, compared to Facebook, YouTube, Twitter or Instagram. Apart from alternative platforms with an international audience, there are several social media with relevance to only specific countries in the analysis. Examples are far-right magazines, 'alternative' and fake news outlets in Austria, RuTube, Jeuxvideo.com and Avenoël in France, Telegram, Discord, Spotify and Tumblr in Germany, Pinterest in Italy or Disqus in Slovenia.

Migration to platforms like VK.com or Gab.ai is often openly advertised on Facebook and Twitter, but also on right-wing websites and blogs. Very often hate speech actors and extremist groups don't give up on the big social media platforms with their far-reaching audience altogether. Instead, they use profiles on different social media networks to reach different target groups. For example, Instagram, a very popular network among young people, is used as an 'eye-catcher' to establish first contact with subtle propaganda. From there, followers of extremist profiles are linked to more explicit and violent content on platforms with a more lenient stance towards hate speech.

Users posting hate speech don't always migrate to already existing social media networks. Sometimes, instead of migrating to already existing social media networks, users open their own website to post hate speech undisturbed. In France, users of the games forum Blabla 18- 25 ans on

Jeuxvideo.com created a new forum called Avenoël when Jeuxvideo started to enforce a stricter policy against hate speech. Another option used by right-wing websites or blogs is the migration into the dark web. The French website Démocratie Participative employed this method to avoid a total shut down after being sentenced in court for illegal content promoting antisemitism, racism, and homophobia and inciting violence.

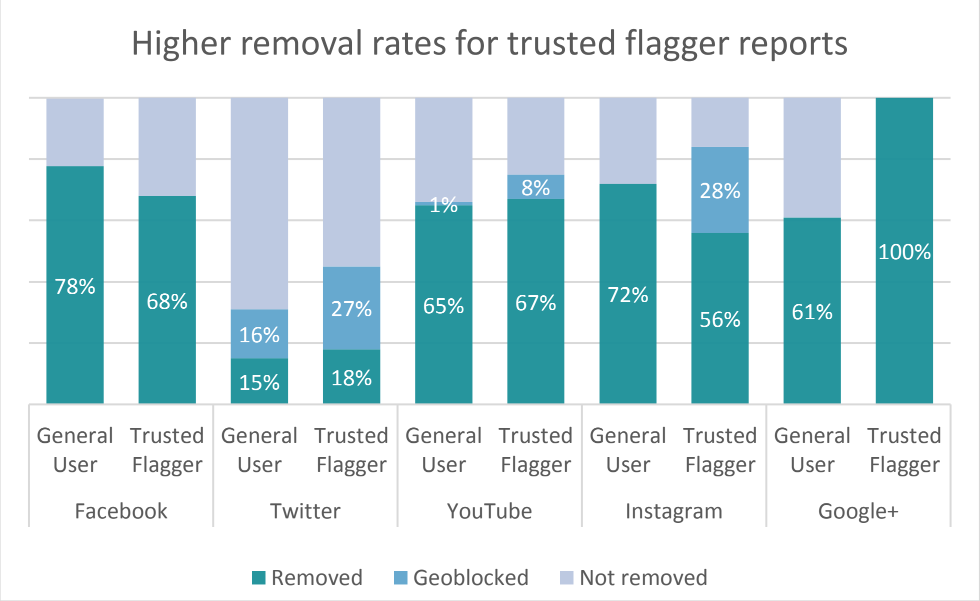
Further analysis is required to keep up with the ever-evolving issue of hate speech on social media. While the relevance of some traditional platforms decreases, new players emerge and new online communities form. Even though some of the platforms analysed currently appear to be relevant only in some countries, the findings in this report can point out potential trends and platforms that should be kept in mind when monitoring hate speech online.

The sCAN Monitoring Exercises

The project partners conducted two monitoring exercises to test the reaction to notifications about hate speech by the IT companies Facebook, Twitter, YouTube, Instagram and Google+. Those IT companies were selected because they have signed the European Commission's Code of Conduct on countering illegal hate speech online. The first monitoring was organised by the European Commission in the period 05.11.2018 – 14.12.2018.

During this period, sCAN partners reported 762 cases of illegal online hate speech to the IT companies Facebook (311 cases), Twitter (190), YouTube (142), Instagram (86), Google+ (23), Dailymotion (8) and Jeuxvidéo (2). In order to test the reaction of IT companies to notifications by their general user base, 755 notifications were sent anonymously through publicly available channels. In a second step, 165 cases that had not been removed after notification as general users were reported again through reporting channels available only for trusted flaggers. Seven cases were reported directly via the partners' trusted flagger channels. Overall, 172 notifications were sent to the IT companies through the trusted flagger channels. The monitored companies took action in 73% of the cases, by either removing (67%) or geo-blocking (6%) the content.

Removal rates differed between the reporting channels used to send the notifications. Overall, the IT companies took action in 62% of content reported via general user channels (58% removal, 4% geo-

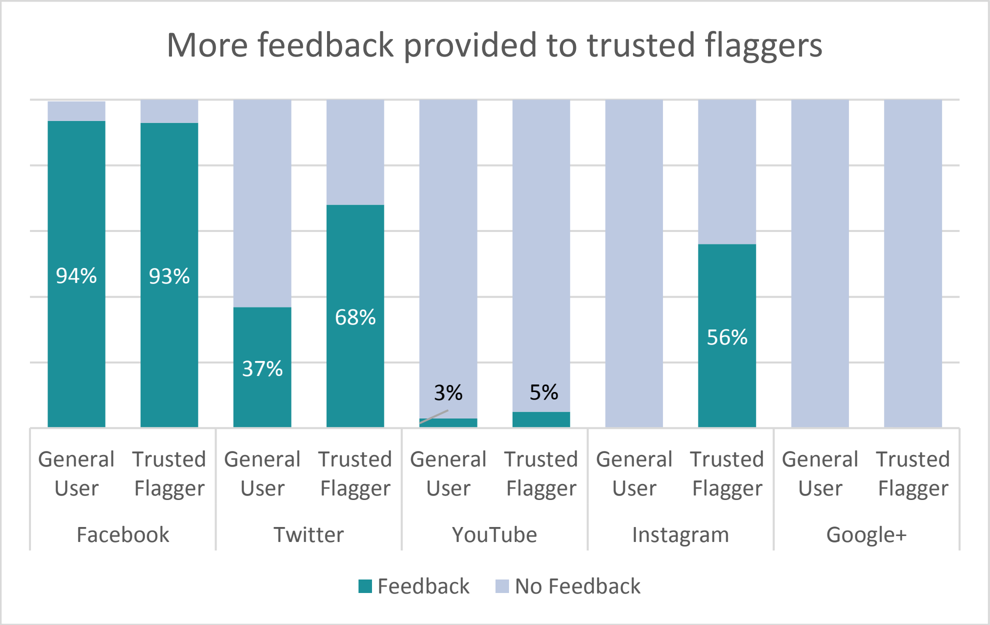


blocking) and in 60% of content reported via trusted flagger channels (42% removal, 16% geo-blocking). Most IT companies reacted more often on notifications sent by trusted flaggers than those sent through reporting channels available to general users of the platforms.

In the Code of Conduct IT companies agree to "review the majority of valid notifications for removal of illegal hate speech in less than 24 hours and remove or disable access to such content, if necessary." 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.

Two of the monitored IT companies removed the majority of content in less than 24 hours after receiving a notification through the channels available for general users: Facebook (76%) and YouTube (58%). Instagram removed 47% of this content in less than 24 hours and Google+ 35%. Twitter removed 12% of content within 24 hours and geo-blocked 13%. When reported through trusted flagger channels, YouTube removed the content in 67% and geo-blocked 8% of the cases in less than 24 hours; Instagram removed 50% and geo-blocked 28%, Twitter removed 17% and geo-blocked 27%, while Facebook removed 32% of the content in this period. Google+ removed none of the content reported by trusted flaggers in less than 24 hours.

Overall, the IT companies provided feedback to 48% of reports through the channels available to general users (46% in less than 24 hours) and to 55% of reports via the trusted reporting channels (45% in less than 24 hours). Facebook was the only IT company systematically providing feedback to all its users, while Twitter and YouTube provided feedback more often to trusted flaggers than to general users. Instagram provided feedback to trusted flaggers only. Google+ did not provide any feedback during the monitoring period. Providing feedback on user notifications is essential to keep users involved and motivated to report illegal content to the companies.



The second monitoring was organised by the sCAN project together with the INACH network in the period 06.05.2019 – 24.06.2019. It largely applied the same methodology as the EC monitoring. The partners used the INACH database as a joint tool for data collection. The results of this monitoring will be published at the End of July.

Trends in online hate speech – partners' experiences

Additionally to the thematic research informing the analytical papers and the monitoring exercises on IT companies' reaction to user notifications, the sCAN partners also keep track of the developments in hate speech in their respective countries.

The Brussels based project partner CEJI identified a general tendency for the normalization of hate speech and the mainstreaming of it on the European level. As politicians shift the public discourse towards more populist approaches, it has become increasingly difficult to fight hate speech. Hate speech is becoming more nuanced and refined and is increasingly used with the purpose of influencing public discourse and the European agenda in general.

In all countries, refugees, Muslims, Jews, Roma and the LGBTIQ community are still the main targets of online hate speech. Intersectionality, i.e. the targeting of several protected characteristics at the same time, can be frequently observed regarding refugees and Muslims. Some partners have identified additional target groups in their countries. In Austria, misogyny is very frequent online. Between September 2017 and September 2018, 83% of all cases of Cyber Bullying reported to ZARA targeted women. In Croatia, the Serb ethnic minority remains a frequent target of cyber hate. In France, anti-Black racism often defames people of color as "uncivilized". Comparisons to monkeys and condemnations of "race mixing" are also common. In Latvia, additional target groups are ethnic Latvians and ethnic Russians. In Italy, xenophobia, anti-migrant hatred and gender based hate speech are the most frequent hate types.

The project partners reported a general trend of the most extreme hate groups online to more strongly utilize the platforms' privacy settings and move their propaganda to closed user groups or networks not accessible to researchers. This trend to move extremist communication to the "dark social"[3] intensifies the problem of so-called echo chambers, where users are only exposed to content reinforcing their already existing hateful views.

An important development in the Czech Republic was the intensified prosecution of hate speech cases. A Romani music celebrity who was targeted with online hate for protesting an award given to a neo-Nazi band has had to sue for redress in different courts all over the country, depending on where the social media user lives. His attorneys have used this example to highlight the need for a more simplified, unified approach to such cases. However, recent public opinion surveys have shown a growing apprehension regarding asylum seekers and Islam, despite the fact that the country has actually only granted asylum to very few people of any religion.

In France, hate speech against Muslims focused on a perceived rise in "Islamic first names". This trend roots in the necessity for racist theorists to come up with concepts to substantiate their claim of a "great replacement" in a country where statistics based on ethnicity or religion are forbidden. Another trend that emerged in France during the past two years is antisemitic rhetoric targeting George Soros. Before that time, antisemitic rhetoric was usually focused on the Rothschild family. Furthermore, since the beginning of 2019 online antigypsyism has multiplied.

In Germany, Islamist online propaganda changed from depictions of graphic violence and calls to join terrorist groups to instructions to commit attacks in their followers' home countries or calls to support

[3] Madrigal, A. C. (2012). *Dark Social: We Have the Whole History of the Web Wrong*. Available at https://www.theatlantic.com/technology/archive/2012/10/dark-social-we-have-the-whole-history-of-the-web-wrong/263523/ (last accessed 24.06.2019).

detained ‘siblings in need’. However, this increasingly subtle propaganda still serves to legitimize violence and glorify the militant jihad. Furthermore, extremist communication in Germany is used to closed channels on messaging apps like Telegram or Whatsapp.

In all analysed countries, the most common tools to disseminate hate speech remain conspiracy theories and fake or biased news, mostly targeting refugees and Muslims. While some of those articles are completely falsified or made up, others combine reporting on current events with biased and unverified information. Growing mistrust in the traditional media landscape leads to a growing popularity of alternative news outlets spreading fake news and conspiracy theories. For example, the French yellow vests movement welcomed the Russian media Russia Today, known for “borderline” fake news or conspiracy theories, while journalists of traditional media faced violent attacks. Furthermore, in the Czech Republic social media influencers increasingly use YouTube videos to spread hateful messages.

During the analysed period, several national and international events have sparked hate speech in the project countries.

Among the international events that triggered hate speech in several countries was the New Zealand terror attack in March 2019. In Slovenia, hateful comments and posts supported the attacker, degrading Muslims and calling for similar acts in Europe or Slovenia. In Germany, Supportive comments were less frequent. Instead right-wing extremists implied that the attack was a ’false flag’ attack perpetrated by a leftist ‘militant eco-activist’ in order to defame the right-wing scene. German Islamists utilized the attack for their own propaganda, claiming it proved that Islam was under attack from ‘the West’. In the Czech Republic, authorities responded immediately and unequivocally to the attack, informing the public that police would be investigating those who expressed approval online for that crime. In several instances, the live video from the shooting was disseminated online. After the Latvian partner reported a link to the video posted by a Latvian Facebook user to the social network, their support contact answered that the video did not violate the Community Standards and would not be removed. This decision was upheld after a request for clarification, even though Facebook officials had stated that the video would be removed from the platform. This shows how important it is for social media to have trained staff moderating content who keep up with current developments and the companies’ decisions regarding recent events. The link to the video has since been removed.

Another event that was met with online hate speech in several countries was the fire at Notre Dame Cathedral in Paris. In Germany and Slovenia, far-right users alleged that the fire was caused by Islamist attackers. In Slovenia, fake news reports about the incident were accompanied by unverified statistical data about the quantity of vandalized churches in ‘multi-cultural France’, inciting hatred and violence towards the Muslim community.

Other hate speech incidents were linked to national or regional events in the partner countries. In Austria, the vice mayor of Vienna, a member of the Austrian Freedom Party, published a picture on Facebook, showing a group of women (wearing headscarves) and children, meeting and having a picknick in a park in Vienna. He added the following comment to the picture: “No far-distance travel picture, but strange impressions from the ‘Türkenschanzpark’ [a park in Vienna]. This is what it looks like on our ‘Viennese’ leisure oases…”. Many of the post commenting on the picture were degrading and hateful towards refugees and Muslims. However, other users reacted with counter speech, pointing out that a park is meant for families spending time outside on a sunny day.

In Croatia, a citizens’ initiative protesting the ratification of the Istanbul Convention against violence against women and domestic violence spread hate speech and fake news against transgender people with aim of their demonization and discrimination in the public. The campaign was also marked by the propaganda of ultraconservative values and the spread of fake news that the Istanbul Convention is promoting a gender ideology.

In the Czech Republic, incidents of antigypsyism online remain very high. In July 2018, Czech Police charged a social media user with inciting hatred in the case of the hateful, racist comments posted beneath a photograph of first-graders in the fall of 2017 because the class was mostly Arab and Romani children. The photo with the names of the pupils and their teachers had been posted to a nationalist website and shared with hateful commentaries by hundreds of social media users. One of the comments was "A grenade would fit in there perfectly…". The arrest of the person posting the comment additionally sparked online hate speech against the police.

In France, the victory of the French national football team at the World Cup 2018 was met with a wave of anti-Black racism online. The hate speech targeted the French football team directly, which was considered as "too Black" for being French and instead presented as an "African" team.

In March 2019, antigypsyist rumours of young women kidnapped on white van by Roma people have sparked a wave of offline violence in France. Violent acts took place in Bobigny and Clichy-sous-Bois, disadvantaged suburbs northeast of Paris. The attackers had armed themselves with baseball bats, knives, and rocks. The rumours appeared on Facebook and Snapchat. Although authorities have dismissed the claims as baseless, online rumours have continued to spread - along with video footage of attacks on van drivers "matching" a supposed description of the alleged kidnapper, across several regions.

In Germany, the conviction of the only survivor of the right-wing terrorist organization National Socialist Underground (NSU), Beate Zschäpe, and NSU supporters in July 2018 sparked a wave of racist hate speech online. While Zschäpe was portrayed as an 'innocent pawn', some of the supporters were styled as heroes. In August 2018, after a fatal stabbing in the German city Chemnitz, refugees were suspected of the crime. In the following days, demonstrations organized by far right groups turned violent, with right-wing extremists attacking people they perceived as migrants in the streets and clashing with police. Right-wing extremists also used these events to incite and influence the discourse online through fake news, manipulations and emotional addresses.

In Italy, a wave of antigypsyist hate speech followed the national media coverage of the confiscation of property of the Casamonica Clan, one of most known criminal groups operating in the periphery of Rome, which has ethnic Roma origins. A police video was published showing the luxurious interior of the confiscated villa. In another incident, famous Italian actress and show girl Asia Argento was targeted with misogynist hate speech. Ms. Argento was among the first promoters of the #MeToo movement, and one of the first women who publicly accused the American producer Harvey Weinstein. Following this, she was herself accused of sexual assault in 2018 by the actor Jimmy Bennett. The alleged episode dates back to 2013. Even though she rejected all the accusations, this led to a skyrocketing increase of misogynist hate speech against her.

In Latvia, discussions around whether the country should join the UN Global Compact for Migration were accompanied by anti-migrant and anti-refugee hate speech in social networks, but also from politicians during parliamentary debates. The Baltic Pride Parade held in Riga in June 2018 triggered an increase in hate speech towards the LGBT persons. In March 2019, a draft law was introduced that would grant foreign students in Latvia the right to full time employment. Right-wing politicians and social media users asserted that non-EU citizens, who would otherwise not be allowed to immigrate to Latvia, would abuse this law to get a work permit without actually studying. Hate speech against ethnic Latvians and Russians was triggered by the amendment of the Education Law by the Parliament introducing new language instruction requirements in Latvian in bilingual (minority) schools. Language issues are particularly sensitive in Latvia and political discourse on the topic triggers online and offline hate speech.

In Slovenia, a right-wing weekly newspaper published a cover depicting a photo-shopped Caucasian woman being groped, un-dressed and attacked by multiple black hands with the tagline ''With migrants, a culture of rape is coming to Slovenia''. One could argue that the cover itself was communicating hate speech and this was rather confirmed when it was posted on various media platforms. It received a backlash from the majority of Slovenian media, who reported on its indecency, while some (right wing) media defended it. Hate speech was significantly present under the news articles of almost all media who wrote about it. In 2018, leaked video-footage of paramilitary gatherings triggered hate speech against migrants online. The footage showed the leader of an extreme right wing political party and former presidential candidate calling for a government coup and a violent “defence” against immigrants. Hate speech re-emerged during the judicial proceedings against the politician, who charged with instigating a violent constitutional change.

Education

It is not enough to simply monitor hate speech online, analyse evolving trends, report and remove illegal content. Media education and counter speech are equally important. In order to build capacity in the civil society and enable users to tackle hate speech wherever they encounter it online, the sCAN partners used the insights gained through the project results and the partners experience in the field to develop online courses and offline trainings on counter-acting hate speech through counter speech, monitoring and moderation of online discussions.

Online course on Hate Speech

The fellow EU project Facing Facts! Online has developed several English language online courses in the thematic area of hate speech, hate crime, bias indicators and advocacy. The sCAN project partners cooperated in translating the course on hate speech into German and French in order to make them accessible to a broader audience.

The course provides background information and equips participants with the knowledge and tools to effectively identify, monitor and counteract hate speech online, e.g. by reporting it to social media platforms or organise counter speech campaigns. The dynamic course employs videos, expert interviews, case studies and online tutoring tailored to the respective national contexts to facilitate the learning process.

The course addresses individual activists, NGOs and public authorities combatting hate speech online, but also to experts and facilitators in the field of media education or citizen education who wish to get a deeper understanding of the ramifications of hate speech and strategies to counter it.

The course is organised in cohorts, which run over a course of six weeks each. The first cohort of the German course was already successfully completed. The next cohorts will start on 08. July 2019 (English course) and 29. August 2019 (German course).

The French course is currently running continuously and is open to registration. Further information on the French course:

Aujourd’hui, sur Internet et notamment sur les réseaux sociaux, les discours racistes et antisémites se multiplient. Les internautes sont de plus en plus confrontés au harcèlement et aux discours haineux sans qu’aucune solution ne se présente à eux pour y faire face. Le projet sCAN et Facing Facts Online vous propose des cours gratuits en ligne afin d’acquérir des méthodes, des outils et des réflexes pour contrer les discours de haine sur internet.

Online Course for Moderators

Building on the general online course on hate speech, the sCAN project also developed an online course on hate speech moderation. The course addresses professionals and individual internet users who supervise online communities hosting discussion boards or comment. It aims to create a better understanding about moderation: the need for it, the tools to support it and the guiding principles of an effective moderation that balances weeding hate speech out of the conversations while also guards freedom of expression.

To maintain healthy conversations online, the course discusses the variety of options for interventions, from removal to counter-speech, and it also encourages participants to create their own moderation policies based on the values they learn to articulate during the course. By the end of the course, participants will be equipped to maintain respect in their online communities, may it be a personal blog's comment section, a YouTube channel or in their professional capacity, e.g. the online forum of a media outlet.

The online course for moderation will be available online on the Facing Facts Online platform in English and French at the end of August 2019.

Advanced Monitoring Training

In addition to online courses, the sCAN project also developed an offline Advanced Monitoring Training. Participants have the possibility to become experts in the field of monitoring hate speech and counteraction, documenting the phenomenon, tackling underreporting, comparing results as well as applying effective human rights reporting. The trainings include interactive sessions on how to recognise hate speech, the importance of monitoring and the art of documentation. An expert trainer from the International Network Against Cyber Hate (INACH) additionally offers a 1-hour morning session in order to give an understanding of how to use the INACH database for documenting cyber hate.

The course addresses individual activists and NGOs who plan to start their own monitoring of online hate speech or seek to professionalise already existing monitoring efforts. The training includes interactive sessions on how to recognise hate speech, the importance of monitoring and methods of documentation. It also covers the European Commission’s Code of Conduct on countering illegal hate speech online and the joint monitoring exercises to evaluate its implementation. To utilize synergies with already existing monitoring facilities, participants also receive training on how to use the INACH database on hate speech for their own monitoring.

Two Trainings on Advanced Monitoring and Countering Online Hatred were already held in Paris (February 2019) and Palermo (June 2019). The trainings were held by Austrian partner ZARA’s training experts and organised by Licra (France) and CESIE (Italy). 36 participants from 10 different countries reflected on the phenomenon in transnational groups, built knowledge and expertise together, gathered best practice examples, and built strong alliances and networks in order to broadly counteract online hatred. Furthermore, the participants had the possibility to dedicate themselves to self-sensibilisation activities (Concentric Circles, Interactive Chat “haters vs. counteractivists”) to understand and differentiate various forms of online hatred, discrimination and cyber mobbing.

Those interested in becoming monitoring experts and contributing to more thorough research on the online hatred phenomenon as well as promoting more visibility and counteraction, can again register for the free sCAN Advanced Monitoring Trainings to be held in Vienna (October 2019) and Brussels (March 2020) in the upcoming months (first-come-frist-served-principle).

Next steps – The sCAN project 2019 – 2020

During the next year, sCAN will continue to monitor and counter hate speech online. The results of the second monitoring, jointly conducted with INACH, will be available at the end of July. Two further monitoring periods will be carried out during the project duration in cooperation with the European Commission and INACH. Furthermore, joint research activities will continue to shed light on hate speech phenomena across Europe. sCAN partners will continue to provide and refine training on different aspects of counter-acting hate both online and offline.

All project results and further information can be found on the project’s blog: www.scan-project.eu

Resources and further reading

sCAN Project resources

sCAN Hate Ontology:

http://scan-project.eu/wp-content/uploads/2018/12/sCAN-D2.3_Hate-Ontology.pdf

sCAN Mapping Study "Countering online hate speech with automated monitoring tools":
http://scan-project.eu/wp-content/uploads/2018/11/SCAN-WP2.1-Mapping-Study.pdf

Analytical Paper "Antigypsyismon the Internet":

http://scan-project.eu/wp-content/uploads/2018/08/Antigypsyism_final-version-1.pdf

Analytical Paper "Beyond the "Big Three" - Alternative platforms for online hate speech":

http://scan-project.eu/wp-content/uploads/2018/08/190529_Beyond_Big3_final.pdf

4th evaluation of the EU Code of Conduct: sCAN project results "Diverging responsiveness on reports by trusted flaggers and general users":

http://scan-project.eu/wp-content/uploads/2018/08/sCAN_monitoring1_fact_sheet_final.pdf

Online Course "Understanding and countering hate speech":

https://www.facingfacts.eu/courses/online-course-on-hate-speech/

Online Course "Hate Speech – Was tun?" (in German):

https://www.facingfacts.eu/courses/hate-speech-was-tun/

Online Course "Combattre les discours de haine sur Internet" (in French):

https://www.facingfactsonline.eu/enrol/index.php?id=22

Online Course "Hate Speech Moderation":

Further information on the course will be available by the End of August 2019.

In English: https://www.facingfactsonline.eu/course/view.php?id=39

In French: https://www.facingfactsonline.eu/course/view.php?id=36

Contact for the offline Advanced Monitoring Training:

Anna-Laura Schreilechner

ZARA – Zivilcourage und Anti-Rassismus-Arbeit
+43 1 929 13 99 - 17
anna.schreilechner@zara.or.at

www.zara.or.at

References

Alliance against Antigypsyism (2017). Antigypsyism - A Reference Paper. p. 5. Available at http://antigypsyism.eu/wp-content/uploads/2017/07/Antigypsyism-reference-paper-16.06.2017.pdf (last accessed 24.06.2019).

International Network Against Cyber Hate (2018). *What is cyber hate*. Available at http://www.inach.net/wp-content/uploads/WHAT-IS-CYBER-HATE-update.pdf (last accessed 24.06.2019).

Madrigal, A. C. (2012). *Dark Social: We Have the Whole History of the Web Wrong*. Available at https://www.theatlantic.com/technology/archive/2012/10/dark-social-we-have-the-whole-history-of-the-web-wrong/263523/ (last accessed 24.06.2019).