CH%# !#ASE

Combatting online HAte Speech by engaging online mEdia

Needs analysis report for online media

Co-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 or European Commission. Neither the European Union nor the granting authority can be held responsible for them.
[Project Number: 101143159]

Combating online HAte Speech by engaging online mEdia (C.HA.S.E.)

CERV-2023-CHAR-LITI-SPEECH
Project reference number: 101143159

Work package 2: Task 2.3, Deliverable 2.4Needs analysis report of existing response practices	
Due date of deliverable	30/10/2024
Submission date	08/11/2024
Organisation name of lead contractor for this document	Symplexis
Author	Thanasis Theofilopoulos
Research team	Paschalia Leventi (Cyprus), Kofi Busumsi (France), Thanasis Theofilopoulos (Greece), Luciano Cortese (Italy).

Version	Date	Summary
1.0	08/11/2024	Development of the final version

The project has been funded with support from the European Commission.
The contents of this publication are the sole responsibility of the authors, and can in no way be taken to reflect the views of the European Union.

Co-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 or European Commission. Neither the European Union nor the granting authority can be held responsible for them.
[Project Number: 101143159]

Contents

Co-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 or European Commission. Neither the European Union nor the granting authority can be held responsible for them.
[Project Number: 101143159]

Foreword

"Combating online hate speech by engaging online media (C.HA.S.E.)" is an EU-funded initiative that tries to address the problem of hate speech[1] based on gender and gender identity that is rampant on the internet and contributes to discrimination and violence. The initiative has been launched in five European countries: Belgium, Cyprus, France, Greece, and Italy.

In the context of the project, the partners in Cyprus (Center for Social Innovation), France (European Center for Human Rights), Greece (Symplexis), and Italy (CESIE European Center of Studies and Initiatives) conducted primary research to identify patterns of online hate speech based on gender/gender identity.[2] They also assessed the needs of online media, identified any beneficial practices they implement, and suggested ways to improve the detection and response to online hate speech comments.[3] The research activities included a combination of qualitative content analysis and visual analysis (online hate speech patterns), as well as focus groups, personal interviews with key stakeholders, professionals, and experts, and an international workshop (to assess needs, gaps, and ways to improve online media). Partners implemented the provisions of a research protocol developed by Symplexis for the project's research purposes. They conducted all research activities involving participants with their informed consent, adhering to the provisions of the EU's General Data Protection Regulation (GDPR).

[1] For the purposes of this publication, "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" (in Council of Europe, Recommendation CM/Rec(2022)16 of the Committee of Ministers to member States on combating hate speech. Available at: https://search.coe.int/cm#{%22CoEIdentifier%22:[%220900001680a67955%22],%22sort%22:[%22CoEValidationDate%20Descending%22]}).

[2] The research results on online hate speech patterns are presented and analyzed in detail in Theofilopoulos, T. (ed.) (2024). *Online hate speech patterns in media platforms' comments sections: Cyprus, France, Greece, Italy*, Project Combating online HAte Speech by engaging online mEdia (C.HA.S.E.).

[3] All results of the research with professionals, experts and stakeholders are presented and analyzed in detail in Theofilopoulos, T. (ed.) (2024). *Online hate speech on the grounds of gender/gender identity: legal framework analysis and mapping of existing response practices*, Project Combating online HAte Speech by engaging online mEdia (C.HA.S.E.).

Co-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 or European Commission. Neither the European Union nor the granting authority can be held responsible for them.
[Project Number: 101143159]

The current publication summarizes the main results of the above-mentioned research activities. Based on these findings, the project will develop a new ICT tool that facilitates the identification of online gender- and gender-identity-based hate speech in real time. Thus, in the near future, an evidence-based tool will be available for online media, empowering them to effectively tackle misogynist[4] and transphobic[5] hate speech comments and contribute to a safer online environment for women and all trans people.

[4] For the purposes of the current publication, misogyny refers to *"hatred or contempt for women"* (in APA Dictionary of Psychology (n.d.) "misogyny. Available at: https://dictionary.apa.org/misogyny). Furthermore, this term *"is derived from the Ancient Greek word "mīsoguníā" which means hatred towards women. Misogyny has taken shape in multiple forms such as male privilege, patriarchy, gender discrimination, sexual harassment, belittling of women, violence against women, and sexual objectification"* (in Srivastava K, Chaudhury S, Bhat PS, Sahu S. Misogyny, feminism, and sexual harassment. Ind Psychiatry J. 2017 Jul-Dec;26(2):111-113. doi: 10.4103/ipj.ipj_32_18. PMID: 30089955; PMCID: PMC6058438.).

[5] For the purposes of the current publication, transphobia is understood as "cultural and personal beliefs, opinions, attitudes and aggressive behaviours based on prejudice, disgust, fear, and/or hatred directed against individuals or groups who do not conform to, or who transgress societal gender expectations and norms. Transphobia particularly affects individuals whose lived gender identity or gender expression differs from the gender role assigned to them at birth, and it manifests itself in various ways, e.g., as direct physical violence, transphobic speech and insulting, discriminatory media coverage, and social exclusion. Transphobia also includes institutionalised forms of discrimination such as criminalisation, pathologisation, or stigmatisation of non-conforming gender identities and gender expressions" (in Yurinova, N. (2023) *Trans Media Guide: A community-informed, inclusive guide for journalists, editors & content creators*. TGEU, p. 38. Available at: https://www.tgeu.org/files/uploads/2023/11/TGEU-Trans-Media-Guide-EN.pdf).

Co-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 or European Commission. Neither the European Union nor the granting authority can be held responsible for them.
[Project Number: 101143159]

Detecting and responding to online hate speech: needs assessment

Online hate speech patterns need to be addressed

The C.HA.S.E. project's research indicates that establishing effective detection and response mechanisms remains a challenging task given the vast extent of the phenomenon in all participating countries as well as the complexity and variety of online hate speech patterns.

During the project's research, many online transphobic hate comments that as such incite hatred, discrimination, and/or violence[6] **against trans people in particular and/or the LGBTQI+ community in general.** The most common patterns— **message/meaning—** identified include

- insulting and promoting hatred by using transphobic slurs and curse words, commonly known in national/social contexts
- promoting and reproducing popular transphobic stereotypes and prejudices, e.g., by treating trans people as mentally ill and trans identities as mental illness
- targeting trans people as a danger to society
- targeting or criticizing human rights defenders and/or public figures because they have advocated for LGBTQI+/trans human rights
- dehumanizing trans people, e.g., comparing them with animals or depicting them as "abnormalities," "non-humans," "garbage," etc.
- encouraging the violation of trans people's human rights, e.g., by asking to put them behind bars or imprison them in mental hospitals
- indirect, encouraging violation of trans people's human rights by glorifying public figures or policies due to their transphobic stances and provisions, respectively
- encouraging trans people to commit suicide or celebrating their death is unacceptable.

[6] See footnote 1.

Co-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 or European Commission. Neither the European Union nor the granting authority can be held responsible for them.
[Project Number: 101143159]

- encouraging violence against trans people, e.g., by justifying it or depicting it as a legitimate act
- inciting acts of violence and humiliation against trans people.

Common **online misogynist hate speech patterns**- in terms of **message/meaning** - inciting hatred, discrimination, and/or violence against women, include

- insulting and promoting hatred by using misogynist slurs and curse words, commonly known in national/social contexts
- promoting and reproducing popular gender stereotypes e.g. regarding the role of women in society and/or how women select their sexual partners or husbands
- promotion of rape culture[7] and victim blaming[8] in cases of gender-based violence incidents, including the trivialization of the phenomenon
- question gender inequality and sexism against women in society
- direct or indirect threats of (sexual) violence
- question or downgrade women's intelligence, skills, and/or accomplishments

In terms of **length or size of the comment**, this may vary from a single word or a single non-textual element (e.g., a GIF, an emoji, a picture) or a single punctuation mark to whole paragraphs of text or combinations of textual and non-textual elements. Furthermore, online hate speech comments on the grounds of gender or gender identity may include **figures of speech, the use of a language similar to leetspeak, or the use of non-textual elements,** thus making detection and reporting a challenging task. Some of them include:

- extended use of metaphors and ironies in a sexist and/or homo/transphobic way that do not necessarily include slurs, curse words that could be easily detected etc. for example, use of reversed terms or slogans to promote sexism (e.g. "toxic matriarchy" instead of

[7] Rape culture refers to a "complex of beliefs that encourages male sexual aggression and supports violence against women" (European Institute for Gender Equality [EIGE] (n.d.). "Rape culture". Available at: https://eige.europa.eu/publications-resources/thesaurus/terms/1314?language_content_entity=en). The term "describes a society where violence is seen as sexy and sexuality as violent" (Ibid). Examples of rape culture may include "slut-shaming", "believing or contributing to rape myths", "victim blaming", "cyber flashing", "image-based abuse" and "misogynistic or homophobic jokes" (Survivor's Network (n.d.). "What is rape culture?". Available at: https://survivorsnetwork.org.uk/resource/what-is-rape-culture/).

[8] "Victim blaming can be defined as someone saying, implying, or treating a person who has experienced harmful or abusive behaviour (such as a survivor of sexual violence) like it was a result of something they did or said, instead of placing the responsibility where it belongs: on the person who harmed them" (Sexual Assault Centre of Edmonton [SACE] (n.d.). "Victim Blaming". Available at: https://www.sace.ca/learn/victim-blaming/). Examples may include phrases like "What did you expect going out dressed like that?", "Why didn't they fight back?", "You shouldn't have gone home with them.", "Why did they get so drunk?" (Ibid).

Co-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 or European Commission. Neither the European Union nor the granting authority can be held responsible for them.
[Project Number: 101143159]

"toxic masculinity") or use of "neutral" words when referring to people's sexual orientation (e.g. calling gay men as "girls") or deliberate misgendering (e.g. using male pronouns when referring to a trans woman).

- repetition of words as well as of punctuation marks (e.g. exclamation marks) and/or non-textual elements such as emojis (e.g. angry, laughing, or vomiting emojis)
- anagramming of homo/transphobic and misogynist slurs and curse words
- removing letters from a slur or curse word, while ensuring that the word is understood by the readers
- replacing letters of a word with other symbols, emoticons or letters, yet the word is still understood by the rest of the online users as a slur, curse word and, generally, a misogynist or transphobic hate speech
- in Greek and Cypriot context, use of "Greeklish" (using Latin letter characters to form Greek words and phrases)
- use of exclamations or single words that are not slurs etc. but are used in insulting, discriminatory ways e.g. "vomit"
- extended use of non-textual elements, namely
    - emoticons depicting objects as sexual symbols or emoticons vomiting to express disgust or humiliate or laughing emoticons to make fun of the people targeted
    - moving GIFS e.g. depicting objects as sexual symbols in a slut-shaming context or depicting people or animals puking or laughing, to humiliate and insult people targeted
    - pictures e.g. pictures of other people making obscene gestures or pictures of public figures known for their anti-LGBTQI+ views
    - memes e.g. with (anagrammed or not) slurs and curse words against women or LGBTQI+ people; memes from movies with actual dialogues that include e.g. homo/transphobic slurs and curse words
    - screenshots from other online content, used in an insulting (misogynist or homo/transphobic) way
    - links to other sources e.g. to YouTube videos with excerpts from movies, and

Co-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 or European Commission. Neither the European Union nor the granting authority can be held responsible for them.
[Project Number: 101143159]

shows, that may include homo/transphobic, misogynist, sexist and other insulting content.

Finally, online media platforms' moderators and managers may find it hard to deal with malicious online behaviors of some users that cannot be reported, as they could hardly be considered violations of terms in online environments. For instance, as the project's research indicated, some online users "tag" the names of other online users in the comments sections, thus "inviting" or "provoking" them to make abusive comments as well. While "tagging" the name of an online user could not be easily treated as a violation of rights, the motivation behind it is a "hidden" or indirect encouragement for such violations.

The next section presents detection, prevention, and/or responding measures—adopted by online media—and relevant gaps identified during the project's research.

Good practices and gaps/needs identified

Participants in the study made reference to online media's combative and preventive tactics. One media stakeholder who took part in the project's research, for instance, mentioned that journalists are encouraged to remain anonymous online and that their organization has policies to respect the LGBTQI+ community. Another media professional who works for a well-known mainstream and popular media company shared his or her thoughts on the steps taken by the organization to address the issue of hate speech by online users, consumers, and visitors under news and posts. According to the participant, their media website already pre-moderates comments using AI technology, which has increased the number of comments received while reducing personal attacks and insults.

In certain instances, the phenomenon is addressed by specialized online media employees rather than AI technology. One research participant, for instance, brought up the function of newspaper staff members who are in charge of spotting and eliminating hate speech and other comments from the newspaper's website. According to a research participant, "managers" of "small [online] groups" or "networks" are in charge of keeping an eye on hate speech comments made online and either removing them or expelling the users who post them. Another participant mentioned how some news media websites have a

Co-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 or European Commission. Neither the European Union nor the granting authority can be held responsible for them.
[Project Number: 101143159]

warning "notice" that states that "any racist, homophobic, transphobic, and comment will be deleted" or a list of online user comments and behaviors that "are prohibited."

Additionally, a few research participants mentioned the reporting options offered by online media platforms. Intentionally abusing online reporting tools to flag accounts with positive content about gender/gender identity issues, however, can result in the removal of posts or accounts or limit their reach, as one research participant observed. Stated differently, the same participant claims that the "community guidelines" and "reporting procedures" offered by social media, online platforms, etc., are not always adequate because they can be used to target accounts and online users who do not have hate speech online content by filing numerous, abusive reports against them.

Generally speaking, online media rarely take any preventive action; when they do, it is usually only after the hate speech comments have been submitted. According to one research participant, online media do not appear to be interested in providing their users or visitors with a "safe space." Some research participants were unaware of any effective strategies or positive actions taken by online media in their nations or overseas to identify, stop, and/or counteract hate speech comments. Even some of the professionals and media stakeholders who participated in the project's research were not entirely certain or aware of the pertinent actions their media or agencies had taken. For instance, a research participant who works for a mainstream media organization admitted that they try to filter comments but expressed uncertainty about how the media handles hate speech. Concerns regarding hate speech and misinformation in online media, specifically with regard to gender and gender identity, were also voiced by another research participant from the media industry. This participant wanted to know if an automatic AI solution was being used to address this problem.

Even worse, some research participants claimed that because online media do nothing to address hate speech or abusive remarks, they appear to accept them. Additionally, hate speech remarks against women and/or the LGBTQI+ community are common and occasionally particularly extreme in certain online media. In order to "cause comments" and "increase their audience," some online media even purposefully employ "provocative titles and themes."

Co-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 or European Commission. Neither the European Union nor the granting authority can be held responsible for them.
[Project Number: 101143159]

There is a notable dearth of information on these issues, and underreporting is also common. There is a general lack of knowledge about the concept of hate speech, or what hate speech is made up of, as some research participants noted. It is clear that one of the police officers who was interviewed was unaware of any hate speech reports made against users of websites or social media accounts by moderators or managers of online media platforms. The majority of these reports are not from the victims themselves, but rather from civil society organizations and/or other internet users. Regulating online platforms and controlling comments about how this is accomplished present additional difficulties, especially on news websites where disparaging remarks are common.

Co-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 or European Commission. Neither the European Union nor the granting authority can be held responsible for them.
[Project Number: 101143159]

Towards better detection and response mechanisms for online media

General recommendations

Research participants were asked to share their thoughts and opinions regarding potential measures, practices, tools, etc. that online media – among other actors – should make use of to better detect, prevent and/or combat online hate speech comments in general and/or on the grounds of gender/gender identity in particular.

Some of the research participants suggested that owners or managers of (media, etc.) channels or (news) websites, blogs, etc. should “ensure the safety” of the internet, by adopting positive detection, prevention, and response measures. Some of the measures suggested include updating the media's code of conduct to include provisions regarding hate speech, implementing improved content moderation practices that employ advanced algorithms and human oversight to identify and swiftly remove hate speech, developing user-friendly reporting tools allowing users to flag harmful content easily, with clear guidelines on what constitutes hate speech.

Moreover, according to research participants, online media should organize relevant information and awareness campaigns (for example, regarding the impact of hate speech and promotion of respectful online behavior), "educate their audience" (for instance, by using "inclusive language" or refraining from using "provocative titles or themes”), publish regular reports detailing their efforts to combat hate speech and the effectiveness of their policies, adopting proactive measures such as warning users about potentially offensive comments before they are posted, collaborate with civil society organizations that specialize in combating discrimination, thus acquiring valuable insights and improving their response strategies. Furthermore, capacity-building activities for media professionals - e.g. on using appropriate terminology or understanding of transgender identities and gender expressions – and establishment of peer support networks within newsrooms for mutual empowerment were also suggested by some of the research participants.

Co-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 or European Commission. Neither the European Union nor the granting authority can be held responsible for them.
[Project Number: 101143159]

To conclude this section, the successful "self-regulation" of media—defined as the regulation of the media "of itself in order to achieve an industry or public policy objective" or "to avoid traditional regulation"—has the potential to have positive consequences. For instance, it has resulted in the establishment of “ethics codes, ombudspersons, and innovative complaints mechanisms that permit news media to remain independent while maintaining high standards”.[9]

Developing a sophisticated ICT tool

Finally, research participants were asked to express their ideas and opinions about the possible features and capabilities of the AI tool that will be created as part of the C.HA.S.E. project and that will make it user-friendly and efficient.

The proposed features and capabilities include recognizing "keywords," "patterns," and "correlations," which helps identify "messages"; understanding the meaning of comments even when they don't contain keywords; and quickly stepping in when a user is about to post a comment, warning them that it will be deleted because it contains hate speech (i.e., when the user clicks "enter" on his or her device, the AI tool will recognize the comment before it is seen by other users); alerting online users about the current legal framework for hate speech; and informing users who have posted hate speech comments about the possible consequences of such remarks on other people (e.g., femicide, suicide, or transphobic hate crimes, giving examples of previous real cases as well).

Additionally, research participants recommended that this AI tool be multilingual, easy to use, and provide clear and understandable information about its operation (how it works) and purpose, even for online users who are unfamiliar with the technology. "Some kind of trusted flagger like audience user" who "has a stake in their community" (of online media users) and is "more likely to feel a responsibility towards the other community members" is another suggestion made by a participant for the inclusion and introduction of the trusted community member/user of online media. According to the same participant, "you get [a] more and loyal user base and you get more trust from users as well if you

[9] Council of Europe (2021 June). CONTENT MODERATION Best practices towards effective legal and procedural frameworks for self-regulatory and co-regulatory mechanisms of content moderation - Guidance Note Adopted by the Steering Committee for Media and Information Society (CDMSI), p. 39. Available at: https://rm.coe.int/content-moderation-en/1680a2cc18

Co-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 or European Commission. Neither the European Union nor the granting authority can be held responsible for them.
[Project Number: 101143159]

promote that and then reward that." Moreover, he/she stated that "there could be levels of use," or, to put it another way, "certain levels of flagging." He/she clarified that a person who violates the rules may, at the very least, initially lose the ability to comment. Additionally, he/she recommended looking into "how those failings might be overcome on a smaller scale" and how large platforms like Facebook's reporting and detection systems fall short.

However, some participants pointed out possible difficulties with using such a tool. One participant added, for instance, that a tool that removes hate speech comments might incite online users to rebel "against the system." In other words, the same participant stated that some users may feel excluded by punitive measures such as an AI tool that automatically finds and removes comments, which could cause them to form alliances and look for "other ways to express themselves." Furthermore, it was argued that an online comment may contain "sexist elements," prejudice, and "stereotypical elements" all at once, making it difficult for an AI tool to handle this "discourse complex." A critical issue in this situation is a careful analysis of the standards for comment removal, whether by a human moderator or an AI tool.

Co-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 or European Commission. Neither the European Union nor the granting authority can be held responsible for them.
[Project Number: 101143159]

General conclusions

Online hate speech on the grounds of gender and gender identity is a widespread phenomenon, taking multiple forms. Hate speech comments include text or non-textual elements or a combination of them, while commentators make use of a variety of figures of speech, making detection and reporting an even more challenging task. While some media have adopted positive measures—mainly focusing on response mechanisms and practices after an online hate speech comment is submitted and detected—most of them do not seem to apply any preventive or counteracting measures at all.

Tackling the phenomenon efficiently requires a multi-level and active involvement of online media, including capacity-building and awareness activities and developing reporting procedures, among others. Given the extent and complexity of the phenomenon that needs to be tackled, an ICT tool to be developed within the framework of the CHA.S.E. project must also fulfill specific requirements—as these have been identified by research participants—to be effective in terms of both user friendliness and detection capabilities.

Co-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 or European Commission. Neither the European Union nor the granting authority can be held responsible for them.
[Project Number: 101143159]