

Artificial Intelligence, Democracy, and Critical Thinking

COUNTERING NEW GENERATION HATE SPEECH

Report on the Comparative Study on Hate Speech
Curated by EuroMed EVE Tunisia and EuroMed EVE Belgium
Developed within the framework of the project :



PROJECT NO. 101185815

ERASMUS-YOUTH-2024-CB



version 1 - mar 2025

Disclaimer

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 the European Commission. Neither the European Union nor the granting authority can be held responsible for them.



SUMMARY

Executive Summary	1
1. Introduction and methodology	5
2. Part A – Documental analysis	9
2.1. Relevant Literature	9
2.2. Legal and policy context	15
2.3. Good practices to tackle online hate speech	27
3. Part B - Field analysis	30
3.1. Country by Country results	30
3.2. Discussion	50
4. Conclusions and final recommendations	53



1. Executive Summary

- The report analyzed the relationship between AI and Hate Speech from both a regulatory and a factual point of view.
- Documental analysis has highlighted that online hate speech has become a major challenge for democracies, especially during political events and elections, driving to the further marginalization of vulnerable individuals and group, suppression of critical voices, fuelling of divisions and violence.
- Digital world – including web 2.0 and Artificial Intelligence (AI) technologies – improves social harm, because of the widens spread of (hate) contents, anonymity, echo chambers, and other factors.
- Hate speech severely affects victims' mental health, requiring psychological support. Young people, highly exposed to social media and online manipulation, are particularly vulnerable, especially those from disadvantaged groups: for example, a study involving over 150,000 young people across 30 countries found that cyberbullying increases the risk of self-harm or suicidal behavior by 2.3 times.
- Online hate speech is variegated across different countries, platforms, political contexts, and country-specific attitudes toward free speech, thus challenging the idea of developing a universal detection system. Hence, findings suggest that context-aware and cross-linguistic approaches are necessary for more effective online hate speech detection. Possible remedies include: 1) Bystanders/users' empowerment; 2) Counter-speech; 3) Digital literacy; 4) Content moderation; 5) Good AI practices, like apps that invite users to reconsider their messages before posting hateful contents, automated detection systems, monitoring tools, feminist and anti-racist online media outlets.

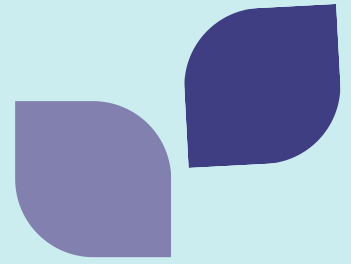
- National and international regulatory solutions play an important role in addressing hate speech. Criminal media regulation and criminal law in general is a primary sector, allowing targeted people to seek redress through the courts. However, the actual impact on reducing hate speech remains under ongoing discussion, especially considering potential side-effects: in particular, activists fear countering hate speech could be instrumentalized by governments to suppress dissent. In the present situation, outstanding legal gaps include: vagueness and ambiguity of normative hate speech definitions; lack of clear digital content guidelines for platforms; limitation and political interference in judicial independence, as well as delays in establishing constitutional judicial review; underreporting by Victims and lack of effective monitoring tools; need for non-criminal and non-repressive legislative solutions.
- On the basis of the identified legal pitfalls, recommendations for governments include: holding and guaranteeing a strong constitutional and legal framework; enacting an effective monitoring system in place, with efficient data collection, mandatory governmental programs, and accessible reporting tools; include non-repressive and non-criminal regulation, such as: promotion of equality, prevention and victim support programs, educational and media literacy programs, intercultural community building programs. Finally, States should invest on human rights driven technological evolution, by developing tools such as AI-driven moderation tools, AI-based monitoring systems, automated reporting systems, or In-App Mental Health Support. Good practices to tackle online hate speech also include Stronger Sanctions for Repeat Offenders, Community-Based Reporting Systems.
- the field research involved a social media content analysis and a study involving a sample from Generation Z (born 1997–2012), ensuring diversity across gender, education level, and linguistic background. Each piece of hate speech content has been evaluated along several dimensions: context, targeted groups or individuals, real or potential impact, and the tone and intention behind the message. A detailed linguistic analysis has further investigated explicit versus implicit forms of expression, and the strategic use of nomination, predication, argumentation, perspectivation, and intensification/mitigation in hate speech.

Following the analysis, researchers have compiled an annotated catalogue of hate speech features to support identification and monitoring efforts. Comparative analyses across partner countries have helped distinguish between universally shared and context-specific characteristics of hate speech.

- Despite the geographic and societal differences across these countries, several cross-cutting themes emerge. Across all six countries, online hate speech is often embedded in everyday digital practices, normalized through humor, memes, satire, and coded language (e.g., jokes, patriot/national identity discourse, secularism...). While the specific groups targeted differ slightly by context, marginalized identities consistently bear the brunt of online hate: ethnic minorities, migrants, women, LGBTQ+ individuals, and people with disabilities. Activists are also targeted, facing heightened aggression.
- The cumulative emotional impact of hate speech is profound. Feelings of alienation, anxiety, distress, and fatigue are frequently reported, particularly among young women, activists, and ethnic or religious minorities. In countries like Tunisia, Italy, and Belgium, this emotional toll leads some youth to disengage from digital platforms. Others, such as participants in Jordan and France, express a heightened sense of polarization and division, particularly during politically sensitive periods. A near-universal frustration with social media platforms emerges in all countries studied. Young people across contexts – especially in Arabic-speaking countries – report moderation systems as ineffective, inconsistent, and opaque. The perceived prioritization of engagement and profit over user safety leads to widespread mistrust. Even when users report harmful content, the lack of response, or arbitrary removals, contributes to a sense of helplessness and institutional complicity.
- Despite the harms experienced, young people are not passive victims of online hate. Across all six countries, they demonstrate resilience, creativity, and a clear desire for structural change. Youth actively engage in counter-speech, educational initiatives, peer support, and advocacy. They suggest innovative solutions, ranging from AI-powered moderation and anonymous reporting tools to community-led monitoring systems and mental health integration within platforms.

- Finally, the findings from these countries underscore that online hate speech is not merely a technological issue but a reflection of broader societal inequalities and fractures. It flourishes where systems fail to protect the vulnerable, and where discourse becomes a battleground for identity, power, and belonging. This scenario calls for urgent reform in platform governance, investment in inclusive digital tools, and the amplification of youth voices as key agents of change. If digital spaces are to become truly inclusive and safe, they must be co-designed with and for the diverse communities that inhabit them.

2. Introduction and methodology



Since 9/11, online hate speech has become a major challenge for democracies, driven by terrorism, political polarization, and the spread of disinformation through web 2.0 and Artificial Intelligence (AI) technologies. Young people, highly exposed to social media and online manipulation, are particularly vulnerable, especially those from disadvantaged groups. Hate speech not only threatens individual well-being but also risks undermining democratic participation and the future of inclusive societies. The purpose of this report is to analyze the matter from both a regulatory and a factual point of view, in order to identify the main needs and recommendations, as well as a possible leeway to introduce a qualitative change in technology and AI, making it more human-centered and attentive to equality and solidarity values. The research of this report, indeed, is functional to the production of an app using AI against hate speech.

For youth organizations, civil society organizations, and institutions, this situation exacerbates a traditional dilemma related to free speech: whether democracy should ‘defend’ itself, by criminalizing discriminatory and anti-democratic language, or trust the ‘marketplace of ideas’, i.e. leave free speech unfettered and assume that hate speech and disinformation will automatically lose influence because they will be defeated in the comparison with good and inclusive arguments. The two extreme positions highlight the need for policies and regulations, at all levels, to balance two interests. On the one hand, repression of hate speech - especially criminal repression - should avoid chilling democratic freedoms of expression and working as a tool for governments to criminalize mere dissent. On the other hand, the ‘free marketplace of ideas’, including the online one, have proven unable to protect human rights, equality, and self-determination, unless accompanied by specific limits for that purpose. For these reasons, the research will devote particular attention to how hate speech is handled in different legislations, which lights and shadows can be identified in criminal tools to counter hate speech, and finally which non-regulatory measures can be enacted, including the development of an anti-discriminatory AI.

This research project explores the nature, dissemination, and impact of hate speech across digital platforms, with a special focus on Generation Z and their online interactions. The objective is to identify the linguistic, social, and contextual dynamics of hate speech and to generate tools for its detection and counter-narration. The research design is multi-method, combining normative document analysis, qualitative interviews and focus groups, a quantitative survey, and in-depth digital content analysis. The documental research involves the examination of a minimum of 10 academic sources, 25 policy documents, 15 legal documents, and 10 pieces of grey literature (e.g., blogs, press articles). Additionally, 10-15 social media contents exhibiting hate speech - such as videos, images, memes, or texts - have been selected from platforms like X, TikTok, Facebook, and Instagram for analysis.

Each piece of hate speech content has been evaluated along several dimensions: context, targeted groups or individuals, real or potential impact, and the tone and intention behind the message. A detailed linguistic analysis has further investigated explicit versus implicit forms of expression, and the strategic use of nomination, predication, argumentation, perspectivation, and intensification/mitigation in hate speech. Following the analysis, researchers have compiled an annotated catalogue of hate speech features to support identification and monitoring efforts. Comparative analyses across partner countries have helped distinguish between universally shared and context-specific characteristics of hate speech.

The project has also involved qualitative data collection through interviews and focus groups with Generation Z participants, ensuring gender balance and diversity in background and digital literacy. A complementary survey targeting a broader Gen Z audience has provided additional insights into their experiences and perceptions of online hate speech. The goal was to understand how young users interact with digital platforms, perceive hate speech, and engage with countermeasures—both personal and institutional. The interviews have been structured around six thematic areas:

1. Social Media Usage: Exploring participants' preferred platforms, usage patterns, and their perceptions of where and why hate speech tends to emerge online.

2. Exposure to Hate Speech: Investigating the types and frequency of hate speech encountered, emotional and practical responses, and perceived trends over time



3. Target Audience and Perpetrators: Examining who is targeted by hate speech, who produces it, and how account visibility or status affects the tone and reach of harmful content.

4. Platform Responses: Assessing users' views on the effectiveness and trustworthiness of social media platforms' moderation, reporting tools, and policy enforcement.



5. User Responses and Solutions: Understanding the extent to which users engage with or propose strategies for addressing hate speech, including education, reporting, activism, or avoidance.

6. Designing Solutions (Mobile App): Gathering ideas on the ideal features of a mobile app dedicated to detecting, reporting, or reducing hate speech.

Participants have been selected from Generation Z (born 1997–2012) and were expected to have a high level of digital literacy. Diversity across gender, education level, and linguistic background is encouraged to ensure a comprehensive dataset. Each interview lasted approximately 30–40 minutes and followed a flexible guideline, allowing participants to share their experiences in a natural conversational flow. Informed consent was a key ethical component of this process. Participants were in fact briefed on the aims, procedures, potential risks, and data confidentiality measures before agreeing to partake. Interviews were voluntary, and participants could withdraw at any time. Data was anonymized unless the participant gave explicit consent to be identified. The findings from these interviews have contributed to a deeper understanding of hate speech dynamics in the digital sphere and will inform the development of both educational resources and digital tools aimed at mitigating online hate. Ultimately, the study has sought to address key research questions on the types of hate speech prevalent on social media, the linguistic strategies that fuel their spread, and the development of effective counter-narratives to disrupt them.

2. Part A – Documental analysis

2.1. Relevant Literature

“Hate speech definition”

The first step of the research requires an interdisciplinary and contextual analysis of hate speech in the different Countries, in order to identify appropriate regulatory remedies^[1]. Different factors can draw discriminatory speech, mainly converging in the attempt to foster a deep hostility towards the “Other”^[2]. Sometimes this form of hatred is built through ‘subtle means’, generating the need to examine the thin line between offensive speech and hate speech: context is particularly relevant, as language can be interpreted differently based on context and intent. Hence, politicians can use coded language or ambiguous phrasing to appeal to certain audiences while avoiding legal repercussions for hate speech^[3]. Concerning religion, for example, the concept of ‘political Islam’ in European counter-extremism policies - through terms like ‘political Islam’ in Austria, ‘legalistic Islamism’ in Germany, and ‘Islamist separatism’ in France - echoes the criminalization of Muslim identity, where the discursive construction of a domesticated ‘national Islam’ - that doesn’t challenge the status quo - is used to justify exclusionary state interventions, marginalize Muslims from political life, enforce compliance, and suppress critical voices^[4].

Online hate speech can fuel radical ideologies, leading to social division and violence^[5], with a profound impact that hate speech can have on societal harmony and the importance of responsible political discourse^[6]. Most importantly, it severely affects victims’ mental health, requiring psychological support^[7]. Youth, in particular, is severely affected by hate speech and cyberbullying^[8]. A study involving over 150,000 young people across 30 countries found that cyberbullying increases the risk of self-harm or suicidal behavior by 2.3 times^[9].

“ Online hate speech ”

Hate speech assumes specific characteristics in the digital world, potentially improving the possibility of social harm. The phenomenon has been studied from different disciplinary angles^[10], including linguistics, communication, psychology, and digital platforms studies, in order to identify personal (e.g., emotions, attitudes), social (e.g., peer influence, group identity), and digital (e.g., platform algorithms, anonymity) factors that contribute to online hate speech^[11]. Online hate speech also has a variegated phenomenology across different countries, platform, political context, and country-specific attitudes toward free speech, thus challenging the idea of developing a universal detection system^[12]. Hence, findings suggest that context-aware and cross-linguistic approaches are necessary for more effective online hate speech detection^[13].

-
- [1] Julien Longhi, Samuel Vernet. Quelle place pour les réseaux sociaux numériques dans la production et la circulation des discours de haine? Réseaux : communication, technologie, société, 2023, N° 241 (5), pp.11-35. <10.3917/res.241.0011>. <hal-04279581>
- [2] Angeliki Monnier. Combattre la haine sur Internet : trois défis à relever. 2019, pp.[En ligne]. <hal-02090674>
- [3] In Discourse and Conflict (Book Chapter)Temmerman, M., & Harder, R. : Different Shades of Hate: The Grey Zone Between Offensive and Discriminatory Language in the Social Media Accounts of Flemish Politicians: https://link.springer.com/chapter/10.1007/978-3-030-76485-2_6
- [4] Hafez, F. (2022). Criminalizing Muslim agency in Europe: The case of ‘political Islam’ in Austria, Germany, and France. French Cultural Studies, 34(3), 313-328. <https://doi-org.libezproxy2.syr.edu/10.1177/09571558221133253> (Original work published 2023)
- [5] Lorenzi Bailly, Nolwenn, et Claudine Moïse, éditeurs. Discours de haine et de radicalisation. ENS Éditions, 2023, <https://doi.org/10.4000/books.enseditions.43765>.
- [6] <https://www.amnesty.org/en/latest/news/2023/03/tunisia-presidents-racist-speech-incites-a-wave-of-violence-against-black-africans/>
- [7] https://www.eeas.europa.eu/sites/default/files/article_de_yosr_jouini.pdf.
- [8] Concerning Jordan, see: Khalil, M. (2024). Cyberbullying and Mental Health Impacts.
- [9] <https://www.khaliltrabelsi.com/tag/cyberbullying/>
- [10] Communications: The European Journal of Communication Research - Vandebosch, H., & Rothmund, T.: Online Hate: A European Communication Perspective : <https://www.degruyter.com/document/doi/10.1515/commun-2024-0097/html>.
- [11] Aggressive Behavior: The Ecology of Online Hate Speech: Mapping Expert Perspectives on the Drivers for Online Hate Perpetration with the Delphi Method : <https://repository.uantwerpen.be/docman/irua/6252dbmotoM75>
- [12] SpringerBriefs in Linguistics - Assimakopoulos, S., Baider, F. H., & Millar, S. : Online Hate Speech in the European Union: A Discourse-Analytic Perspective <https://link.springer.com/book/10.1007/978-3-319-72604-5>
- [13] CLIPS Technical Report Series- De Smedt, T., Jaki, S., Kotzé, E., Saoud, L., Gwózdź, M., De Pauw, G., & Daelemans, W. : Multilingual Cross-domain Perspectives on Online Hate Speech <https://repository.uantwerpen.be/docman/irua/e092ae/156589.pdf>

Despite different contexts, common features of the Internet can be identified as factors fueling hate speech: the possibility of having anonymous profiles[14], the tendency of social media to produce echo chambers as key drivers of hate speech propagation[15].

In different parts of the Mediterranean area, studies highlight increased polarization during political events and elections[16], with severe effects on youth engagement[17]. Minority groups, refugees, and political opponents[18] are targeted; religious intolerance on social media platforms has a relevant societal impact, too[19]. Gender is also a major cause of online discrimination, particularly targeting activists, journalists, and politicians[20]. Young women are also affected by sexist stereotypes spread online[21].

“Remedies”

1 Bystanders/users' empowerment

Users' empowerment stands out as a main source to counter hate speech, hence raising the need to explore the conditions that foster it. Studies have explored the factors that influence whether bystanders intervene when witnessing online hate speech or cyberbullying: personal affection, digital literacy, and support from others stands out as relevant factors, even if many users fear retaliation or believe interventions are ineffective. This suggests that platforms should encourage counter-speech and provide tools to support bystanders in reporting hate speech[22].

[14] Al-Sharif, R. (2022). Digital Identity and Hate Speech Dynamics

[15] Abu-Zaid, M. (2024). Social Media Narratives and Extremism; Salim, T. (2023). Social Media Algorithms and Echo Chambers.

[16] Mansour, F. (2023). Political Discourse and Hate Speech Online

[17] Since the 2011 uprising, Tunisia's digital space has become a battleground for politically charged debates, often including harmful content and gender-based attacks: Al-Momani, H. (2023). Online Hate Speech and Youth in Jordan

[18] IAMCR Conference Paper- Braeckman, A.: Hate Speech on the Internet: Case Study of Belgium and the Netherlands
<https://biblio.ugent.be/publication/374903>

[19] Omar, L. (2024). Religious Intolerance and Digital Narratives

[20] https://www.eeas.europa.eu/sites/default/files/article_de_yosr_jouini.pdf

[21] Yasin, A. (2023). Gendered Hate Speech on TikTok

[22] Behaviour & Information Technology-Rudnicki, K., Vandebosch, H., Voué, P., & Poels, K. : Systematic Review of Determinants and Consequences of Bystander Interventions in Online Hate and Cyberbullying Among Adults
https://repository.uantwerpen.be/docman/irua/fda381/main_document_.pdf

2 Counter-speech

Research has underscored the importance of empowering users to actively participate in online discussions and counteract hate speech, highlighting the role of community engagement in fostering a respectful online environment^[23]. While the remedy might seem ineffective at first glance, there are factors that have proven to increase the impact of counter-speech, such as humor, warning of consequences, and empathy. The findings on racist hate speech revealed that empathy-based counter-speech had a small but measurable effect in reducing hate speech, whereas humor and warnings of consequences were less effective^[24]. The role of influencers and community leaders can also prove relevant to the purpose^[25].

3 Digital literacy

Digital literacy programs to mitigate negative impacts^[26].

4 Content moderation

Social media platforms like Facebook and Twitter that monitor and control hateful behavior, using a mix of automated detection and user-reported content^[27]. These tools - both human moderators and AI-based detection systems - have shown limitations, particularly when dealing with coded language, humor, and implicit hate speech. Additionally, there are risks that private and non-public institutions can generate^[28], particularly with regards to possible chilling effects on hate speech. These circumstances highlight the need for coupling these remedies with different regulatory tools^[29].

^[23] Ascone, Laura/Placzynka, Karolina/Vincent, Chloé (2023). Counter narratives against antisemitic hate speech. A comparative study of French and British online comments sections. English version for Wissen schafft Demokratie. Schwerpunkt Netzkulturen und Plattformpolitiken, Band 14. Available exclusively online at <https://www.idz-jena.de/wsdet/wsd14-12>

^[24] Proceedings of the National Academy of Sciences: Dominik Hangartner, Gloria Gennaro, Sary Alasiri, Nicholas Bahrach, Alexandra Bornhoft - Empathy-Based Counter-Speech Reduces Racist Hate Speech: Evidence from a Field Experiment <https://www.pnas.org/doi/10.1073/pnas.2116310118>

^[25] Zain, K. (2023). Counter-Narratives Against Hate Speech

^[26] Haddad, N. (2024). Youth Digital Literacy and Hate Speech; IAMCR Conference Paper- Braeckman, A.: Hate Speech on the Internet: Case Study of Belgium and the Netherlands <https://biblio.ugent.be/publication/374903>

^[27] Angeliki Monnier. Combattre la haine sur Internet : trois défis à relever. 2019, pp.[En ligne]. <hal-02090674>

^[28] Emmanuel Netter. À la recherche du juste rôle des plateformes exploitant des réseaux sociaux dans la lutte contre la "haine en ligne".. Centre de droit privé et de sciences criminelles d'Amiens. Les plateformes d'intermédiation numérique, 2022

^[29] Exploring AI Governance in the MENA Region : Gaps, Efforts, and Initiatives

5 Good AI practices

Technology can also power original responses to hate speech. For example:

- the "Rethink" app by Trisha Prabhu aims to prevent online hate by prompting users to reconsider their messages before posting^[30];
- techniques have also been experimented for the automatic detection of jihadist hate speech in Belgium and beyond. This technology uses machine learning to analyze patterns in online extremist rhetoric. In that sense, lexical and contextual analysis methods can improve automated detection systems for online hate speech, to face the challenge of differentiating between ideological expression and direct incitement to violence^[31];
- Hatemeter Platform is an automated tool that collects and analyzes large dataset from social media, identifies potential hate messages in real-time, visualizes data through various graphs, and offers suggestions for crafting counter-narratives. This platform assists social science researchers in exploring extensive data and supports NGOs in monitoring social media, training personnel, and accelerating the production of counter-narratives to combat hate speech^[32];
- digital feminist media outlets also have a role in promoting gender equality and combat hate speech^[33].

[30] <https://www.khaliltrabelsi.com/tag/cyberbullying/>

[31] CLiPS Technical Report Series- De Smedt, T., De Pauw, G., & Van Ostaeyen, P. Automatic Detection of Online Jihadist Hate Speech; <https://repository.uantwerpen.be/docman/irua/eeeddb/149408.pdf>

[32] Mario Laurent. Project Hatemeter: helping NGOs and Social Science researchers to analyze and prevent anti-Muslim hate speech on social media. *Procedia Computer Science*, 2020, 176, pp.2143- 2153. 10.1016/j.procs.2020.09.251 . hal-03493729

[33] Fagundes-Ausani, Mariana. 2025. "Combating Anti-Feminism and Building Strategies for Gender Equality: The World of Digital Feminist Media Activism in Brazil and France". *Politikon: The IAPSS Journal of Political Science* 59 (1). Online:33-62. <https://doi.org/10.22151/politikon.12025.2>.

6 Regulatory solutions

Criminal media regulation^[34] and criminal law in general is a primary sector where hate speech is addressed at European and international level despite the challenges in defining hate speech and the unavoidable role of national courts in interpreting and applying relevant laws^[35]. Some Mediterranean examples show that such legislation can support targeted people to seek redress through the courts^[36], but the actual impact on reducing hate speech remains under ongoing discussion^[37]. Additionally, repressive approaches can also pose challenges, as activists fear countering hate speech could be used to suppress dissent^[38].

International treaties can also play a role, even if their domestic implementation may remain fragmented. Hence, there is the need for cohesive national strategies and improved coordination among institutions to effectively combat online exploitation^[39]. In that sense, alternative policy approaches can be explored, based on better education and awareness for magistrates and the public, as well as transparency, platform accountability, and better user reporting mechanisms^[40].

[34] Evan Raschel. "Panorama de droit pénal de la presse (septembre 2019 - mai 2020)", Lexbase Pénal, 2020, n° 28. Lexbase Pénal, 2020, n° 28. <hal-02903868>

[35] Les discours de haine, le droit pénal et la Cour européenne des droits de l'homme, Thomas Hochmann, revue de droit et libertés fondamentales, RDLF 2024 chron. n°05

[36] <https://merip.org/2021/08/the-limits-of-confronting-racial-discrimination-in-tunisia-with-law-50/>

[37] Marwa Jabrane, "Online Freedom of Expression and the Legal Regulation of New Media"; <https://www.arab-reform.net/publication/tunisias-law-against-racial-discrimination-the-mixed-results-of-a-pioneering-legislation/>

[38] https://www.eeas.europa.eu/sites/default/files/article_de_yosr_jouini.pdf

[39] Dorra Loudhaief Midani et al., "The Legal, Policy, Institutional and Programmatic Context for Protecting Children Against Online Sexual Exploitation in Tunisia

https://www.researchgate.net/publication/382642512_The_legal_policy_institutional_and_programmatic_context_for_protecting_children_against_online_sexual_exploitation_in_Tunisia

[40] European Union Agency for Fundamental Rights: Online Content Moderation: Current Challenges in Detecting Hate Speech: <https://fra.europa.eu/en/publication/2023/online-content-moderation>

2.2 .Legal and policy context

The international framework for hate speech

To provide a unified framework for the United Nations to address the issue globally, the UN Strategy and Plan of Action on Hate Speech (2020) defines hate speech as «any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor». According to this definition, hate speech can be conveyed through any form of expression (including images, cartoons, memes, etc., both online and offline) and should be distinguished from hate crimes, as well as other acts of discrimination, because it is based on a form of expression. Additionally, hate speech is defined as such when it refers to a person identity factor laid down in the UN definition, but also other factors that are not explicitly recognized. Hence, online and offline bullying of children and youth can also be considered as hate speech especially when it is based on such factors. Communication that refers to two or more identity factors is common and can increase the harm suffered by its targets. United Nations field presences' responses to hate speech should be calibrated according to the level of severity, assessed on the basis of (a) the social and political context; (b) the status of the speaker; (c) the intention of the speaker; (d) the content and form of the speech; (e) the extent of its dissemination; and (f) the likelihood of harm, including imminence (the “six-part test” or the “Rabat threshold test”). Only incitement to discrimination, hostility or violence that meets all six criteria should be criminalized, while less severe forms of incitement or hate speech should attract civil or administrative law based restrictions, or public policy responses. Countermeasures should consider that hate speech can be lawful or unlawful.

The UN Strategy identifies 3 levels:

1) The severest forms of hate speech **are prohibited under international law**. Such expressions include, most notably: (a) “direct and public incitement to commit genocide”, as defined by international criminal law (art. 3 (c) of the Convention on the Prevention and Punishment of the Crime of Genocide; and arts. 6 and 25 (3) (e) of the Rome Statute of the International Criminal Court); and (b) “any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence”, as defined in article 20 (2) of the International Covenant on Civil and Political Rights (ICCPR). In addition, article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination imposes a duty upon States to criminalize “all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another colour or ethnic origin”. According to international standards: the terms “hatred” and “hostility” should be understood to refer to intense and irrational emotions of opprobrium, enmity and detestation towards the target group; Advocacy should be understood as requiring an intention to promote hatred publicly towards the target group; “Incitement” refers to statements which create an imminent risk of discrimination, hostility or violence against persons belonging to targeted groups.

2) Certain forms of hate speech **may be prohibited under international law**, even if they do not reach the above-mentioned threshold of incitement, in specific circumstances (article 19, International Covenant on Civil and Political Rights). Such limitations need to: (a) be provided by law; (b) pursue a legitimate aim, such as the respect of the rights of others, including the right to equality and non-discrimination, or the protection of public order; and (c) be necessary in a democratic society and proportionate (the “three-part test”).

3) The least severe forms of hate speech **must not be subject to legal restrictions under international law**. This category includes offensive, shocking or disturbing expressions – including the condoning or denial of historical events, blasphemous speech, and disinformation – when they do not reach the threshold provided in Art. 20 ICCPR. In these cases, non-legal responses might be needed, including a “plurality of policies, practices and measures nurturing social consciousness, tolerance and understanding change and public discussion”.

The table below – taken from the UN Strategy and Action Plan – shows a framework for assessing the severity of hate speech, derived from the Rabat Plan of Action: the Rabat threshold test.

CRITERIA OF SEVERITY	INDICATORS	QUESTIONS TO ASK
1. Context	The legal, political, social and economic context	<ul style="list-style-type: none"> • Is there an ongoing conflict or are there incidents of violence against the targeted group? • Does the law recognize the targeted individual's or group's identity? • Is there any anti-discrimination legislation and is it aligned with international human rights norms and standards? • How does the media report on the targeted group, if at all? • Is the media independent? • Are there upcoming elections? • What is the role of identity politics in electoral campaigns? • Are there any challengers to the hate speech? If so, who are they?
2. Speaker	The position or status of the speaker in society and their authority or influence over their audience	<ul style="list-style-type: none"> • Does the speaker have power or influence in society? • Are they a national leader, politician, public official, religious or faith leader, or social media influencer? • What is their reputation and standing in society? • What is their relationship with targeted groups?

3. Intent	The state of mind of the speaker	<ul style="list-style-type: none"> • Did the speaker intend to engage in advocacy of hatred against an individual or group on the basis of a protected characteristic? • Was there a triangular relationship, that is, did the speaker intend to incite the audience against the target group? (in the case of incitement only) • Was the speaker merely negligent or reckless in their expression? • Was the speaker's communication in poor taste or showing a lack of judgment?
4. Content and form	Nature and style of the expression	<ul style="list-style-type: none"> • To what extent was the speech provocative and direct? • What was the form, style and nature of arguments deployed in the speech? • Was there any balance of arguments deployed in the expression? • Was the expression in the public interest? • Was it artistic or academic expression?
5. Extent and magnitude of the expression	Reach of the expression	<ul style="list-style-type: none"> • How public was the expression when it was made? • How widely was the expression disseminated? • How large was the audience that was exposed to the expression? • Was the expression disseminated offline and/or online?
6. Likelihood, including imminence	Degree of risk of harm	<ul style="list-style-type: none"> • Was there a reasonable probability that the speaker's communication would succeed in inciting actual action by the audience against the target group? (in the case of incitement only) • Was there a reasonable probability that harm would result from the expression? (e.g. physical and/or psychological harm to an individual or group, or harm to social cohesion) • Would the harm have affected particular individuals within the targeted group (e.g. women, children or youth) more than others? • Would the harm have a different impact upon women and men?

Finally, the UN Strategy contains 13 commitments:

1. Monitoring and analysing hate speech
2. Addressing root causes, drivers and actors of hate speech
3. Engaging and supporting the victims of hate speech
4. Convening relevant actors
5. Engaging with new and traditional media
6. Using technology
7. Using education as a tool for addressing and countering hate speech
8. Fostering peaceful, inclusive and just societies to address the root causes and drivers of hate speech
9. Engaging in advocacy
10. Developing guidance for external communications
11. Leveraging partnerships
12. Building the skills of United Nations staff
13. Supporting Member States

Other international sources relevant for online hate speech are:

- International Convention on the Elimination of All Forms of Racial Discrimination (1966)
- International Covenant on Civil and Political Rights (1967)
- UNESCO (2022). Policy Framework on Digital Citizenship
- OECD (2023). Guidelines for Responsible Digital Communication
- UN Women (2022). Gender-Based Online Hate Speech Report
- UN Convention on Digital Rights (2023)
- 2024 UNESCO's Ethical AI Guidelines

The European regulatory landscape

Council of Europe, Recommendation of the Committee of Ministers to member States on combating hate speech (Adopted by the Committee of Ministers on 20 May 2022) – following Recommendation No. R (97) 20 of the Committee of Ministers to Member States on "hate speech" (Adopted by the Committee of Ministers on 30 October 1997 at the 607th meeting of the Ministers' Deputies) defines hate speech 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». This definition should also take into account the following factors and the interplay between them: the content of the expression; the political and social context at the time of the expression; the intent of the speaker; the speaker's role and status in society; how the expression is disseminated or amplified; the capacity of the expression to lead to harmful consequences, including the imminence of such consequences; the nature and size of the audience, and the characteristics of the targeted group. The Recommendation contains guidelines to protect human rights and fundamental freedoms in the digital environment, against hate speech. Measures include: criminal law measures, civil and administrative law remedies, as well as other countermeasures for offensive or harmful types of expression which are not sufficiently severe or evident to be legitimately restricted under the European Convention on Human Rights but nevertheless call for alternative responses. Specific measures should be defined by States in the online environment, requiring Internet intermediaries to fulfil their duties and responsibilities not to make accessible or disseminate hate speech that is prohibited under criminal, civil or administrative law. Important elements include: rapid processing of reports of hate speech; removing hate speech without delay; respecting privacy and data-protection requirements; securing evidence relating to hate speech prohibited under criminal law; reporting cases of such criminal hate speech to the authorities; transmitting to the law-enforcement services, on the basis of an order issued by the competent authority, evidence relating to criminal hate speech; ensure transparent, proportionate, and human rights compliant content moderation; ensure human monitoring of AI content moderation tools. Non-criminal measures should also include monitoring of hate speech and support to targeted people

The EU Code of conduct on countering illegal hate speech online (2016) includes commitments from the platforms regarding reporting, analysis without delay and removal of hate content. In addition, IT companies commit to conducting among their users education and awareness of the types of unauthorized content according to the rules and guidelines prepared by them for the user community. The companies are involved in the development and promotion of independent alternative narratives, new ideas and initiatives, and supporting educational programs that encourage critical thinking. Monitoring of this tool, however, has identified some critical issues related to both the delicate balance between the removal of material reported as “inappropriate” and the preservation of the right of expression, as well as the possibility that some subjects migrate to platforms less regulated. Its assessment (2019) shows an increased rate of content removals, but not always highlighting of a context-specific monitoring and acting.

EU High Level Group on combating racism, Xenophobia and other forms of intolerance, Key guiding principles on encouraging reporting of hate crime the role of law enforcement and relevant authorities (2021), analyses barriers – such as social-cultural biases and difficulty in accessing supporting services – as well as promoting structures and processes, including training of public officers and cooperation with civil society. EU Anti-Racism Action Plan 2021 “The EU Anti-Racism Action Plan 2020-2025 aims to strengthen legislation and enforcement against racism, involve affected communities in policymaking, promote equal opportunities in employment, education, and housing, combat racial profiling, and improve law enforcement practices”.

The European Declaration on Digital Rights and Principles for the Digital Decade (2022) provides a framework to ensure a safe and fair digital environment and a human centered digital transformation, based on values of solidarity and inclusion, freedom of choice. The Declaration also includes specific provisions on the protection and empowerment of children and young people in the digital environment, in order to ensure that they are «empowered to make safe and informed choices and express their creativity in the digital environment». Prevention and protection from online crimes should be ensured, but also training and literacy opportunities, promotion of positive experiences, safeguard against all children and young people against harmful and illegal content, exploitation, manipulation and abuse online. Finally, the Declaration promotes the involvement of children and young people in the development of digital policies that concern them.



Next Generation EU (Regulation (EU) 2020/2094 and Regulation (EU) 2021/241) focuses on three main priorities: accelerating the green transition, enhancing competitiveness through innovation and digital transformation, and strengthening social and territorial cohesion.

Digital Services Act^[41] (Regulation (EU) 2022/2065) (DSA) aims to create a safer online environment for consumers and companies in the European Union (EU), by: defining clear responsibilities for online platforms and social media; deal with illegal content and products, hate speech and disinformation; achieve greater transparency with better reporting and oversight; and encourage innovation, growth and competitiveness in the EU's internal market. The regulation introduces responsibilities and a system of accountability and transparency for platforms, including special rules for Very Large Online Platforms (VLOPs) – used by more than 10% of the 450 million consumers in the EU – and Very Large Online Search Engines (VLOSEs) – used by more than 10% of the 450 million consumers in the EU. The Regulation provides transparency rules for platforms and aims to ensure that effective notification mechanisms are in place, in order for people to be able to easily flag illegal content or products, hate speech and disinformation. Additionally, it establishes a list of ‘trusted flaggers’, with priority on such notifications, thus allowing independent and qualified control of online contents. Dispute resolution mechanisms are included to challenge content-moderation decisions and seek redress, through a dispute mechanism or judicial redress.

Digital Markets Act (Regulation (EU) 2022/1925) (DMA) establishes a set of clearly defined objective criteria to qualify a large online platform as a “gatekeeper” and ensures that they behave in a fair way online and leave room for contestability. Hence, a clear set of obligations is provided for these platforms in order to keep the market fair for users and competitors. These rules, by challenging the dominance of platforms, might influence the way disinformation is spread through ‘echo chambers’ and ‘filter bubbles’, as these ones are largely nourished by the prevalence of big players that dominate information flows based on data-driven business models. Indeed, the more contestable the market is, the more it will be possible for users to choose different and more human-rights-compliant platforms.

[41] Communications: The European Journal of Communication Research - Vandebosch, H., & Rothmund, T.: Online Hate: A European Communication Perspective : <https://www.degruyter.com/document/doi/10.1515/commun-2024-0097/html>.

European Commission, Joint communication to the European Parliament and the Council, No place for hate: a Europe united against hatred, Brussels, 6.12.2023, JOIN(2023) 51 final provides specific measures for the online world, such as: extending networks of trusted flaggers on social media working on all grounds of incitement to violence and hatred, to support and anticipate enforcement of the Digital Services Act, including a dedicated network of trusted flaggers specialised on antisemitism.

European Media Freedom Act (Regulation (EU) 2024/1083) (EMFA) aligns with existing national efforts to uphold media freedom, by introducing EU-wide standards to further protect media independence and transparency. Additionally, rules are provided to guarantee the position of media – including small and independent media – to be protected in case of removal or deindexation of contents by online platforms.

EU High Level Group on combating hate speech and hate crime, Key guiding principles on cooperation between law enforcement authorities and civil society organisations (2023) lays down a series of guidelines and principles for law enforcement authorities to cooperate with civil society organisations (CSOs). Such principles include the need to establish a culture of trust and transparency in order to safeguard the CSOs' independence together with the sustainability of the cooperation. Cooperation can include third-party reporting through CSOs, data collection and monitoring, victims' support, and other activities of collaboration and training.

Country by Country results landscape

The analysis of the situation in each Country has sought to identify:

- A) Legal and policy gaps;**
- B) good policy practices;**
- C) recommendations**

A) Legal gaps include:

- *Vague Definitions and Ambiguity.* Sometimes legislative provisions lack clear, precise, and up-to-date rules to identify hate speech and differentiate it from other crimes or from free speech in general. This gap can lead to: 1) inconsistent application of the law on personal and territorial bases; 2) ineffective enforcement, which hampers the protection of groups like women, LGBTQ+ people, and religious minorities; 3) especially in the framework of counterterrorism legislation, there is a potential abuse and misuse in favour of majority and privileged groups, in situations of persisting inequalities;
- *lack of clear digital content guidelines for platforms.* No institutionalized AI-based content moderation at the national level. Inadequate reporting mechanisms for victims;
- some Countries experience *limitation and political interference in judicial independence*, as well as delays in establishing constitutional judicial review. In other Countries, judicial review is formally guaranteed, but there is insufficient enforcement of court decisions, particularly concerning human rights violations;
- *underreporting by Victims.* Many people do not report hate speech due to fear, lack of trust in authorities, or not knowing where to turn for help. This means a significant amount of hate speech goes unnoticed and unpunished;
- *criminal legislation needs to be completed by other kinds of legislation* addressing: economic inequalities leading to marginalisation; insufficient data collection and right to access to public documentation, complaints or rulings related to hate speech; lack of attention for prevention.

B) Good policy practices

1. *Constitutional framework:* Constitutional protections for human rights, ensuring equality and prohibiting discrimination, help curb hate speech.
2. *Clear Laws:* Create specific laws that define and punish hate speech.
3. *Strong and Consistent Enforcement:* Train authorities to enforce these laws fairly and consistently, ensuring appropriate monitoring of speech, particularly online. For example, the Pharos Platform (run by the Ministry of the Interior) monitors and investigates online hate speech.
4. *Judicial review:* Some courts have convicted individuals for incitement to hatred or violence, setting examples and raising public awareness.

5. *Platform Rules:* Make social media platforms remove hateful content and ban repeat offenders.

6. *Education:* Teach solidarity, equality, and respect in schools and communities. Enact media literacy programs

7. *Support Victims:* Provide help and protection to those targeted by hate speech.

8. *Public Awareness Campaigns:* Governments and civil society organizations run campaigns to promote understanding and counter hateful ideas.

9. *International collaboration:* EU, UN, and other bodies/agreements promote human rights initiatives, which sometimes include hate speech components

C) Recommendations for governments include:

1) A strong constitutional and legal framework

- strengthen constitutional guarantees for human rights and equality, including for judicial independence and against corruption;
- introduce a clear legal definition of hate speech in national law, in line with international human rights standards (e.g., ICCPR);
- ensure counterterrorism laws respect civil liberties;
- promote stronger Regulations for Social Media Companies at all State and international levels. For example, DSA requires tech companies to take down hate speech more quickly and be more transparent about their moderation policies. Governmental legislation can be complemented with self-regulation and co-regulation processes encouraging social media platforms to cooperate with authorities for localized content regulation. For example, the Code of Conduct on Disinformation has encouraged online platforms to adopt the strengthened guidelines on countering illegal hate speech online, which are now integrated into the Digital Services Act.

2) An effective monitoring system:

- improve data collection and monitoring by implementing a system to gather comprehensive data on hate speech and hate crime incidents. This data should be disaggregated by motivation bias and the victims' characteristics, including disability, gender, and age.
- implement mandatory programs focused on discrimination for all law enforcement and criminal justice officials.

- make Reporting Easier & More Accessible. Many people don't know where or how to report hate speech. A more user-friendly reporting system, including multilingual hotlines and digital tools, could encourage more victims and witnesses to come forward.

3) Include non repressive and non criminal regulation, such as:

- the enforcement of labor laws and the promotion of economic, social, and territorial inclusion inclusion;
- educational and media literacy programs, equipping individuals with the skills to recognize and counteract such content;
- increase Support for Victims. Hate speech can have serious psychological effects on victims. More counseling and legal assistance services should be available, especially for marginalized groups. This can include: Establish Helplines: Set up telephone and online support services to assist victims of cyberbullying, offering advice and psychological support; Community Support Groups: Create support groups for individuals affected by psychophobia, fostering the exchange of experiences and solidarity. Monitoring Trends: Implement mechanisms to monitor the evolution of cyberbullying and adjust policies accordingly;
- invest in publicly funding Community Programs. Increase financial aid and cooperation with NGOs that work on anti-racism, inclusion, and community dialogue. Create a table of "bottom-up" propositions of regulations in this field.

4) Invest on human rights driven technological evolution, by developing tools such as:

- AI-driven moderation tools to detect and flag hate speech on social media and websites in real-time, tailored to different languages and dialects
- AI-based monitoring systems that could help authorities track and remove hate speech online more effectively
- Automated Reporting: Implement systems to automatically report harmful content to authorities or platform moderators.

2.3 . Good practices to tackle online hate speech

As hate speech continues to pose a significant challenge across social media platforms, young users have outlined several practical and innovative solutions to address the issue and create a safer online environment. These solutions focus on enhancing moderation, increasing user accountability, and providing emotional support for those impacted by online abuse. Below are the key proposals gathered from participants:



AI-Powered Moderation Tools

A significant number of users emphasized the importance of leveraging Artificial Intelligence (AI) to combat hate speech. The proposal centers around the implementation of AI-powered moderation tools that can automatically detect and filter harmful content in real time. These technologies would help minimize the reach of hate speech by intervening before such content spreads across platforms. By using advanced algorithms, AI systems can flag and remove offensive material more quickly, reducing users' exposure to hateful content.



Stronger Sanctions for Repeat Offenders

Participants also suggested that more stringent penalties should be enforced for users who consistently engage in hate speech. The implementation of stronger sanctions, such as temporary suspensions, permanent bans, or even legal actions in extreme cases, would serve as a deterrent to repeat offenders. By holding users accountable for their actions, social media platforms could foster a more respectful online culture and reduce the prevalence of harmful speech.



Community-Based Reporting Systems

There was a widespread call for the introduction of community-based reporting systems that would increase transparency throughout the reporting process. Many users expressed frustration with the lack of visibility into how reports are handled, which often leaves them feeling that their complaints are ignored. A transparent reporting system would allow users to track the progress of their reports and understand the actions taken against offenders. This would encourage more users to report hate speech, knowing that their concerns are being addressed in a fair and timely manner.



In-App Mental Health Support

Given the emotional toll that exposure to hate speech can take on individuals, many users proposed that social media platforms offer in-app mental health resources. These resources could include access to licensed therapists, AI-powered chatbots offering emotional support, or peer support groups. By providing immediate psychological help, platforms could offer users a safe space to process the emotional impact of online abuse, potentially helping to prevent long-term mental health consequences.



Educational Initiatives

To combat hate speech at its root, many participants advocated for in-platform educational initiatives aimed at promoting media literacy and encouraging respectful behavior. These initiatives could take the form of awareness campaigns, educational posts, pop-up prompts, or videos that inform users about the negative effects of hate speech on individuals and communities. Such efforts could also encourage users to engage in constructive and positive conversations, helping to shift the overall tone of online interactions. Some participants proposed the inclusion of interactive games within the app. These games would allow users to practice responding to hate speech scenarios, helping them learn how to use constructive counter-speech and recognize manipulative or provocative tactics.





Envisioned Anti-Hate Speech Apps

In addition to improving existing social media platforms, some participants envisioned the development of dedicated anti-hate speech apps that could serve as comprehensive tools for both combating online abuse and supporting its victims. These apps would integrate various features designed to detect, respond to, and prevent hate speech, while using advanced AI algorithms to monitor and remove hate speech as it occurs. This would ensure a more immediate response to harmful content, preventing its spread across social media. The app could also include therapeutic chatbots, stress management resources, and peer support forums to help users manage the psychological impact of exposure to hate speech, thus fostering a supportive digital community dedicated to promoting respectful communication. Users could share resources, engage in positive discussions, and support one another in combating hate speech.

3. Part B - Field analysis

3.1. Country by Country results

This social media content analysis presents a detailed investigation of various online examples that illustrate how racism, sexism, homophobia, ableism, xenophobia, and religious intolerance are manifested and normalized across digital platforms. By critically examining the impact and intention behind each piece of content, it was possible to gain insights into the broader societal consequences of online hate speech and the nuances of tone and rhetorical strategy used to disguise or intensify it.

Belgium

In the case of Belgium, the analyzed examples include memes, social media posts, news incidents, and user-generated comments, each revealing distinct patterns of discrimination and dehumanization. These patterns often rely on humor, sarcasm, or “cultural tradition” as a smokescreen for spreading harmful narratives, reinforcing stereotypes, and alienating marginalized groups. Several cases highlight how racism and sexism are trivialized or dismissed as jokes, while others expose more openly aggressive, inflammatory, or provocative tones designed to provoke fear, rally extremist ideologies, or undermine trust in institutions. Key themes emerging from the analysis include: the normalization of prejudice through humor and satire; dehumanization practices via imagery and language (e.g., through animal comparisons, emoji use); institutional complicity or inaction, as seen in responses from schools, media, or social platforms; intersectional discrimination, where race, gender, sexuality, and disability overlap in the marginalization of individuals; resistance and advocacy, including counter-narratives that reclaim agency and promote inclusivity.

The interviews reveal that young people from diverse backgrounds are frequently exposed to hate speech on social media platforms, particularly on Twitter and Facebook. The most commonly targeted groups include Muslims, refugees, Black communities, LGBTQ+ individuals, women - especially those in public or male-dominated spaces - and ethnic minorities. Hate speech often spikes during elections, political events, and global crises, reinforcing harmful stereotypes and fueling division. Participants consistently reported frustration with platform responses. Reporting mechanisms are widely seen as ineffective, slow, and inconsistent - many hate-filled posts remain online, while others are removed without clear criteria. This lack of accountability undermines trust in social media companies and leaves victims feeling unsupported. The impact of hate speech is both emotional and psychological. Many interviewees described feeling unwelcome or alienated in their communities. To cope, they turn to friends and family, avoid online arguments, engage in self-care activities, or temporarily withdraw from social media. When asked about solutions, participants emphasized the importance of education, especially around diversity, digital awareness, and critical thinking. They also proposed stronger platform policies, more transparent moderation, and better tools for users to detect and report hate speech. Some imagined apps or browser extensions that use AI to flag harmful content, provide context, and offer support for those affected.

The focus group participants, primarily using Instagram, TikTok, Twitter, and Reddit, engage with these platforms for various reasons, including sharing personal content, news, memes, and participating in discussions. Instagram and TikTok are popular for sharing creative content like photography, dance videos, and fashion, while Twitter and Reddit are more focused on news, memes, and debates. Despite the active engagement, participants express concern about the prevalence of hate speech, particularly in comment sections and public forums. Hate speech is a frequent issue for many participants, with insults often targeting personal identities, including race, gender, and hobbies. Participants, such as Youssef and Aya, have encountered racist remarks and body shaming. Hate speech is especially prevalent during political events, cultural debates, or discussions on sensitive topics like immigration and climate change. Many participants report that hate speech has become more frequent and intense over time, largely attributed to social media algorithms and the anonymity of users.

In response to hate speech, participants employ various strategies. These include reporting comments, blocking users, avoiding toxic spaces, and countering hate speech with positive content or constructive debates. However, many participants express frustration with the lack of effective moderation by platforms, as reported hate speech often goes unaddressed. Regarding platform trust, participants generally feel that social media companies are ineffective in managing hate speech. There is a widespread sense that platforms prioritize engagement over user safety, which allows hate speech to spread unchecked. Many participants call for better moderation tools, faster response times, and more transparency in how platforms handle reports. They further suggest implementing stricter platform policies, improving moderation tools, and launching educational campaigns to promote digital literacy. Additionally, participants advocate for user involvement in the moderation process, such as community-driven reporting systems. Some others suggest the creation of real-time hate speech detection apps or features, along with support resources like counseling or legal assistance for victims.

Jordan

In the case of Jordan, the main concerns that emerged from the social media content analyzed include the escalation of religious intolerance, gender-based discrimination, xenophobia, and the suppression of critical speech. Several high-profile cases and events reveal the impact of online hate speech on individuals and communities, as well as the broader societal implications of a lack of effective moderation and enforcement. An emblematic reported example is the assassination of journalist Nahed Hattar after posting a controversial cartoon online, which underscores the growing religious intolerance in Jordan. Hattar's death, following death threats from extremists, highlights the peril faced by secular voices in a polarized environment. The incident exacerbated the hostility between religious conservatives and critics of religious norms. Another thought-provoking example was the public outcry against Jordan's cybercrime law, which criminalized online speech, and raised significant concerns about freedom of expression. Journalists and activists warned that this law suppresses critical speech and increases self-censorship, further emboldening pro-establishment voices and silencing dissent. Furthermore, hate speech against Syrian refugees has become widespread in the country, with online calls to close borders and blame refugees for economic and social issues.

This rhetoric has fostered increased discrimination, hostility, and exclusionary attitudes within public discourse, further alienating marginalized groups. Women's rights activists and feminists in Jordan have also faced significant online harassment, particularly related to the controversial Article 308, which allowed rapists to marry victims to avoid punishment. This has led to threats and defamation campaigns targeting those calling for stronger anti-harassment laws, with conservative voices resisting gender equality reforms. On the other hand, disinformation campaigns targeting minority groups and human rights organizations have resulted in damaged reputations and a spread of divisive narratives. These campaigns, often driven by conspiracy theorists and far-right voices, exacerbate societal divisions and hinder efforts for peace and inclusion.

Young people in Jordan actively use social media platforms such as Instagram, TikTok, and Twitter for creative expression, social connection, and political engagement. However, many of them encounter hate speech and online discrimination related to gender, religion, ethnicity, and political views. These negative experiences often intensify during trending events, political debates, or cultural discussions. Common targets include women, religious and ethnic minorities, political activists, and mental health advocates. Perpetrators are frequently anonymous users or accounts with large followings, and platform responses are widely viewed as ineffective or inconsistent. For example, Sondos, a 19-year-old digital design student, has experienced targeted online hate, especially as a woman. Despite reporting harmful content, she sees little impact, often turning to friends for emotional support. Her suggestion is to create safer online spaces through educational tools and anonymous reporting features. Azzam, a 21-year-old university student, prefers Twitter for political discourse but notices increased polarization and hostility during election periods. He highlights that even verified accounts contribute to hate speech. He copes by using humor and actively reporting abusive content, calling for better tools like sentiment analysis and real-time moderation. Aya, an art student, reports rising religious intolerance online, particularly around religious events. Despite her efforts to report hate speech, she finds Instagram's response slow and ineffective. She seeks more inclusive digital environments where users can safely share experiences, including through anonymous forums and peer support groups.

Farah, a freelance designer, uses Instagram and TikTok to showcase her work but faces ethnic discrimination, especially during global political crises. She takes breaks from social media to protect her mental well-being and advocates for AI-powered content moderation and stronger filtering tools. Mohammad, a graduate student in psychology and mental health advocate, experiences cyberbullying and stigma during his awareness campaigns. He finds current reporting systems opaque and calls for more transparency. To combat online hate, he combines educational responses with self-care practices and suggests integrating mental health support into social platforms.

This focus group conducted in Jordan explores the experiences of 50 young adults in Jordan (ages 18–30) from diverse backgrounds, including students, professionals, and activists, regarding hate speech on social media. Participants frequently use platforms such as Facebook, Twitter (X), TikTok, Instagram, and YouTube. An overwhelming 90% reported encountering hate speech online, with the most common forms being xenophobia (particularly targeting Syrian refugees), religious intolerance, and misogyny. Women, minorities, refugees, and political activists are the primary targets, while perpetrators are often anonymous users, ideological groups, political actors, or influencers. Platform moderation is widely viewed as ineffective, especially due to limitations in Arabic-language AI moderation. Most users rely on personal strategies: 60% ignore hate speech, 25% respond with counter-speech, 10% report it, and only 5% pursue legal action. Additionally, many adopt protective behaviors such as using VPNs or anonymous accounts. Overall, the findings reveal that hate speech is widespread, often unaddressed, and leads users to rely on self-protection rather than institutional mechanisms. There is a clear need for more robust moderation systems, Arabic-language AI tools, and accessible support channels.

Hate speech in Tunisia's digital landscape is both pervasive and multifaceted, reflecting deeper social inequalities and prejudices. It often escalates through emotionally charged, derogatory, and exclusionary language that incites fear, justifies discrimination, and reinforces systematic marginalization. The lack of effective moderation on platforms allows this speech to persist, indicating a pressing need for stronger content regulation, public awareness campaigns, and inclusive digital practices. By way of example, Black Tunisians and migrants are frequently excluded from the national identity and portrayed as inferior or threatening. This discourse relies on labeling ("foreigners"), harmful predictions ("uneducated", "criminals"), and emotionally charged arguments ("stealing jobs") to dehumanize and marginalize these groups. At the same time, refugees, particularly from sub-Saharan Africa, are vilified as criminals and threats to public safety and national resources. These portrayals foster xenophobia and justify exclusionary or hostile attitudes. Women also face dismissive and aggressive rhetoric questioning their capabilities in professional and public roles. Common narratives portray them as irrational or overly emotional, reinforcing patriarchal norms and limiting gender equality. Anti-Islamic content on social media uses extreme language - such as likening Islam to a "disease" or "cancer" - to justify fear and violence. This hate speech is often framed from a secular or majority perspective, escalating societal tension. People with disabilities are also frequently described as "burdens" or "unproductive", reinforcing stereotypes that contribute to their social and economic marginalization. The language used is often patronizing or openly hostile, portraying disability as a personal and national liability. Furthermore, online interactions around mental health reveal both supportive and harmful tendencies. While some users offer empathy and solidarity, others express dismissive or stigmatizing views, contributing to an unsafe digital environment for mental health discourse.

The interviews conducted in Tunisia capture insights from young individuals aged 19 to 29 regarding their experiences with hate speech on various social media platforms, including Instagram, Facebook, Twitter, TikTok, YouTube, and Discord.

The participants highlighted regular exposure to multiple forms of hate speech, ranging from racism, sexism, and political hate to religious intolerance, body shaming, and discrimination against LGBTQ+ communities. Hate speech was often encountered in comments, memes, videos, and direct messages, with the majority of incidents originating from anonymous accounts or public profiles. Participants noted that hate speech often intensifies during periods of political instability, such as elections or when controversial news events unfold. These events appear to act as catalysts, amplifying the prevalence of hateful rhetoric across social media platforms. Public figures, influencers, and identity-based groups (such as women, ethnic minorities, and LGBTQ+ individuals) were consistently reported as primary targets of hate speech. Hate speech related to an individual's physical appearance, including harmful comments about body size, shape, or appearance was also frequently reported in the interviews. This content often leads to emotional fatigue, anxiety, or frustration for those exposed to it. Witnessing hate speech or being exposed to it regularly had a significant emotional impact on users. Many reported experiencing distress, with some choosing to distance themselves from social media platforms to protect their mental health. While some participants actively engaged in discussions to combat hate speech, others reported feeling overwhelmed by the pervasive negativity. Despite reporting mechanisms, many users expressed frustration with the limited effectiveness of existing moderation tools. Moderation tools across the platforms were criticized for being inconsistent and ineffective. Participants gave moderation effectiveness scores ranging from 0% to 40%, highlighting a significant gap in user satisfaction. Key issues identified were: the enforcement of community guidelines often lacked uniformity, with some users facing penalties for minor infractions while others engaged in harmful behavior without consequence; slow response times leading to users feeling discouraged from engaging with reporting systems. In light of this, the general sentiment was that reporting systems were insufficient in addressing hate speech effectively, leading to diminished trust in the platforms' ability to manage harmful content.

Within the French context, an analysis of various social media posts reveals recurring patterns of representation that target immigrants, Arabs, and Muslims in ways that promote division, dehumanization, and social exclusion. A common feature across these posts is the construction of a clear in-group versus out-group dynamic. The in-group, which is typically framed as French citizens, integrated immigrants, or individuals upholding societal norms, is positioned as the victim of an alleged threat posed by the out-group, which includes immigrants (particularly Black Africans, Maghrebis, Romanians, and Afghans), Arab individuals, and Muslims. For instance, immigrants are frequently nominated using derogatory or reductionist labels such as "illegal aliens" and are predicated as disruptive, criminal, or parasitic. These portrayals are often supported by anecdotal examples or symbolic representations, such as color-coded maps that visually associate high immigrant density with undesirable living conditions. Such visual rhetoric subtly reinforces the message that immigrant presence correlates with social decline. Similarly, Arab individuals are targeted through posts that emphasize ethnic distinctions, sometimes under the guise of humor. Terms like "Arabe mental" not only carry pejorative connotations but also perpetuate the notion that Arabs are inherently irrational or problematic. These portrayals reinforce racial stereotypes and contribute to broader social narratives that depict Arabs as outsiders who fail to integrate into French society. Muslims, in particular, are subjected to some of the most intensified forms of stigmatization. Posts and videos that label Muslims as violent, antisemitic, or inherently dangerous use emotionally charged language, such as "baby killer" or direct calls for their removal, to invoke fear and justify Islamophobic attitudes. These narratives often frame Islam itself as a threat, thereby justifying discriminatory behavior as a form of societal self-defense. However, not all posts follow this pattern of exclusion. Some aim to raise awareness about the psychological effects of racial discrimination, highlighting concepts like *charge raciale* (racial burden). These narratives shift the focus from the alleged shortcomings of marginalized communities to the systemic forces that oppress them. They emphasize the daily pressures faced by individuals who must constantly modify their behavior to navigate a society structured by racial and cultural biases. What emerges from this analysis is a consistent pattern in which minority groups are blamed for societal problems, their experiences are dismissed, and their identities are reduced to stereotypes.

Insights from interviews with several young adults aged 22 to 27 reveal a consistent pattern of exposure to hateful content, dissatisfaction with platform moderation, and a desire for more effective intervention. Participants reported varying degrees of exposure to hate speech. While some encountered it occasionally, others described frequent and persistent exposure, particularly women who were often targeted with sexist content. Hate speech was also commonly directed at identity-based groups, especially along racial and religious lines. Some individuals noted that specific people were directly targeted, highlighting the personal and often aggressive nature of such content. The most commonly used platforms among interviewees were Instagram, Facebook, TikTok, and LinkedIn. While these platforms are mainly used for entertainment, socializing, and professional networking, they also serve as hubs where hate speech circulates. Despite the popularity of these platforms, participants expressed a lack of trust in their effectiveness in managing hate speech. Many felt that current reporting systems and moderation tools were either ineffective or too slow to act meaningfully. A recurring theme in all interviews was the role of anonymous accounts in spreading hate speech. The perceived impunity that anonymity offers was seen as a significant barrier to accountability. Users felt that anonymity emboldened individuals to post harmful content without fear of consequences. Participants shared that they commonly report or block accounts that spread hate speech. Some also engage in counter-speech, attempting to challenge hateful narratives or educate others. However, these personal efforts are often met with frustration due to a lack of visible platform action. Only a few interviewees had participated in broader anti-hate campaigns, suggesting a potential gap in awareness or motivation to engage in organized digital activism. Almost all participants proposed improvements to address the issue more effectively. For example, stricter moderation policies and real-time content detection, the use of artificial intelligence to automatically flag and remove hate speech and stronger consequences, such as temporary bans or permanent removal of repeat offenders. They also envisioned educational tools within platforms as important resources to promote awareness and encourage respectful discourse.

Regarding the focus group, most users reported spending time on platforms like Instagram, Twitter, TikTok, Facebook, and LinkedIn. These platforms are chosen not only for their popularity but for their ease of communication, diverse content, and ability to follow trends and pop culture. Users engage primarily through liking, commenting, sharing, and messaging. Yet this everyday interaction is often marred by the presence of hate speech. A significant number of participants, especially women and minority users, shared that they regularly come across sexist, racist, xenophobic, or fatphobic content. This hate typically appears in the form of comments on public posts, memes, news articles, or viral content. Twitter and TikTok, in particular, were identified as platforms where hate speech is most prevalent. Respondents cited that anonymity and lack of accountability empower perpetrators to express themselves with aggression and hostility. Many users described feeling more exposed to hate during periods of political or social tension, such as elections or public debates on gender equality. One participant pointed to a surge in misogynistic content around France's elections, while another noticed spikes in hateful rhetoric during International Women's Day. These events often trigger waves of online abuse targeted at women, minorities, or politically active users. Despite the frequency of hate speech, users expressed frustration with current platform responses. Reporting tools are often seen as slow, ineffective, or inconsistent. Very few respondents believe their individual actions, like reporting or blocking, actually lead to meaningful change. Some choose to disengage by taking breaks from social media, while others seek emotional support from friends and family. When imagining solutions, users proposed thoughtful and practical ideas for combating hate speech online. Many suggested real-time content moderation using artificial intelligence, more efficient and accessible reporting features, and larger moderation teams to respond to flagged content quickly. Some emphasized the importance of community-based moderation systems, drawing inspiration from platforms like Reddit, where niche spaces with clear rules often foster safer environments. Several participants also proposed preventative features, such as ID-based registration to limit anonymity, keyword filters, and educational tools to promote awareness and empathy. A recurring theme in their feedback was the need for platforms to take stronger responsibility, not just by removing hateful content but by actively cultivating inclusive and respectful digital spaces.



Morocco

In Morocco, the social media analysis shows the growing prevalence of hate speech targeting minority groups as well as discriminatory depictions of Moroccans in international media. The analysis focuses on textual content across social media and public discourse, highlighting key patterns of racism, religious intolerance, and sexism. More specifically, digital platforms have seemingly seen a rise in racist commentary targeting Sub-Saharan African migrants, especially in major urban centers. These narratives construct migrants as "outsiders", fostering hostility and social division. The impact includes heightened fear, social isolation, and increased susceptibility to discrimination and violence. Moreover, religious minorities, such as Christian converts and Jewish communities, seem to be frequently targeted with intolerant rhetoric on social media and in public forums. This results in a pervasive sense of vulnerability and discourages public expression of faith. The discourse often frames these groups as "unfaithful" or "foreign", reinforcing their marginalization. Female journalists in Morocco, particularly those reporting on controversial social or political issues, also emerge as a group facing demeaning and hostile online abuse. This hate speech undermines their credibility, contributes to self-censorship, and has adverse psychological and professional effects. Common arguments question women's right to speak publicly or hold authoritative roles in media. Official reviews and reports have begun addressing racism and xenophobia in Morocco. These documents cite hate speech against Sub-Saharan migrants, refugees, and ethnic minorities (e.g., some Berber communities) as major concerns. Institutional discourse emphasizes the need for legal reform and stricter enforcement to combat hate-based discrimination. The analysis also revealed Islamophobic media coverage in international contexts. Two notable tweets criticized German media for allegedly comparing Morocco's national football team to ISIS following the 2022 World Cup. Such framing intensifies negative stereotypes and fosters hostility toward Moroccan and Arab communities. The discourse employs sarcastic and accusatory tones to expose perceived European Islamophobia and racism. The analyzed content reveals consistent patterns of exclusionary discourse, often rooted in racial, religious, or gender-based bias. Addressing these issues requires both platform-level moderation and policy-level interventions that promote inclusion, accountability, and media responsibility.



Interviewees reported encountering a wide range of hate speech, particularly targeting minorities, such as Black people, Amazigh people, Muslims, and religious minorities like non-Muslim converts. The most common forms of hate speech included racism, Islamophobia, sexism, and discrimination based on ethnicity. Hate speech often spikes during controversial events, such as polarizing social issues, elections, or international controversies, where public sentiments are more divided. Most interviewees expressed dissatisfaction with the effectiveness of social media platforms in handling hate speech. While there were some improvements, such as AI-driven content moderation and reporting tools, the platforms' responses were perceived as inconsistent and often subjective. Many participants found the reporting tools unreliable or slow, with little impact from individual reports. This has led to frustration and a lack of confidence in the platforms' ability to address the problem. Reactions to encountering hate speech varied. Some interviewees, such as Azzeddine and Nour-Eddine, tend to ignore hate speech but occasionally report it when necessary. Others, like Firdaous, reported blocking accounts when exposed to repeated hate speech. However, few participants actively engage in combating hate speech beyond reporting it, largely due to feelings of frustration and helplessness. A common coping strategy among participants was to take short breaks from social media to alleviate emotional distress. A recurring recommendation was to focus on education, both for users and platform developers. Interviewees suggested implementing in-app alerts, educational modules, and context labels for hate speech, which could help users better understand the impact of their actions. Several participants, including Azzeddine and Yasser, advocated for enhanced AI-powered moderation tools to better identify and remove hate speech in real time. They also emphasized the need for platforms to use smarter, context-sensitive moderation systems. Many interviewees called for more robust and enforceable regulations to combat hate speech, including clearer policies, faster response times to reports, and stricter penalties for offenders. The emotional toll of witnessing hate speech, particularly for vulnerable groups, was highlighted as a significant concern.

The findings from the focus groups further reveal a significant concern about the prevalence of hate speech on social media platforms and the perceived inefficiency of current moderation strategies. Participants, who included university students, young professionals, and others, frequently encountered hate speech across various platforms such as Facebook, WhatsApp, Instagram, TikTok, and Snapchat. This hate speech predominantly targeted minorities, women, political activists, and specific ethnic groups, such as Amazigh people. It was often propagated by anonymous accounts, making it difficult for participants to hold perpetrators accountable. Notably, political events or controversial periods led to spikes in hate speech, further intensifying participants' exposure to harmful content. A common frustration among participants was the inconsistency of platform moderation. Many expressed skepticism about the effectiveness of reporting tools, often feeling that their reports were ignored or handled too slowly. This lack of follow-through created a general sense of distrust in the platforms' ability to address and mitigate harmful content. Despite some improvements in platform policies, harmful content continued to persist, with participants noting that responses were frequently inadequate. In terms of user responses, while some participants chose to confront hate speech, the majority opted to ignore it due to frustration, fear of retaliation, or a sense of helplessness. This inaction was compounded by the feeling that efforts to report or challenge hate speech were futile, especially when no visible changes followed. Additionally, several participants expressed concerns about retaliation, particularly when engaging with sensitive or controversial topics, such as political activism or social justice issues. Regarding potential improvements, participants suggested the integration of AI-driven real-time moderation as a way to address harmful content more swiftly and effectively. Many called for better reporting tools and the implementation of real-time alerts, alongside educational resources to help users recognize and report hate speech. There was also a strong desire for clearer and more transparent community guidelines, as well as reminders to help users understand the consequences of engaging in harmful behavior. In addition to these suggestions, participants emphasized the need for stricter platform policies, faster response times to reported content, and more consistent enforcement of these policies.

The social media analysis shows hate speech events in Italy targeting individuals and groups through social media platforms such as X (formerly Twitter) and Facebook. The posts and comment threads analyzed reveal a pattern of hate-driven discourse mobilized primarily by political polarization, sexism, xenophobia, and ableism, often aimed at delegitimizing political opponents, activists, or minority groups. The targets involved were specifically university students, female public figures, NGOs, political representatives from left-wing parties, and migrants, while the perpetrators political figures, media outlets, and online users aligned with far-right or conservative ideologies. Sexist discourses, for example, targeted women in politics (e.g., Stefania Proietti) and public figures (e.g., Elodie) with objectifying, misogynistic language and visuals. Xenophobic discourses on the other hand were directed at NGOs (e.g., Mediterranea Saving Humans) assisting asylum seekers, portraying them as criminal collaborators in a modern slave trade. Explicit and implicit incitement to hatred and violence emerged from the analysis but also group polarization, where “in-groups” (right-wing conservatives, nationalists) are framed as morally superior and defenders of national values, while “out-groups” (leftists, migrants, progressive women) are demonized. These contents reinforce gender-based discrimination and undermine female political participation, normalize xenophobic narratives and criminalize humanitarian efforts, deepen ideological divides and encourage hostility toward political opposition, and exploit cultural trauma (e.g., fascist resistance) for political gain. These cases demonstrate a disturbing normalization of hate speech within mainstream digital discourse in Italy, often weaponized by or amplified through political and media figures. This environment not only targets vulnerable groups but also distorts public opinion, undermines democratic engagement, and erodes civil discourse.

Interviews were conducted with eight young individuals (ages 21-27) to investigate their experiences with online hate speech (HS), their perceptions of existing platform responses, and their ideas for potential solutions. Participants, primarily students, regularly use platforms like Instagram, TikTok, YouTube, and Discord.

While some have been direct targets of online hate, such as sexism, cyberbullying, and body shaming, others have primarily witnessed it, often directed at marginalized groups including women, people with disabilities, LGBTQ+ communities, and ethnic or religious minorities. Across all interviews, there is a shared lack of trust in current platform reporting systems. Most participants feel that reporting mechanisms are either ignored or inconsistently applied, with little to no follow-up. As a result, many choose to ignore hate speech entirely or respond with limited actions such as reporting or commenting. A few attempt to open discussions when the content is particularly personal or offensive. Despite varying levels of optimism, all participants offered ideas for new tools to address hate speech. Common themes include the creation of a standalone app that allows users to report incidents and receive support from real human experts, not bots. Features such as 24/7 expert assistance, educational resources, hate speech detection across platforms, and progressive sanctions for offenders were frequently proposed. Some imagined forum-style spaces for moderated discussion, while others envisioned keyword-based detection systems that could identify and flag harmful content. While a few interviewees believe in the power of education to combat hate speech, others argue that education alone is not enough. Several expressed doubt about the success of awareness apps, citing low user engagement and increasing desensitization to online hate.

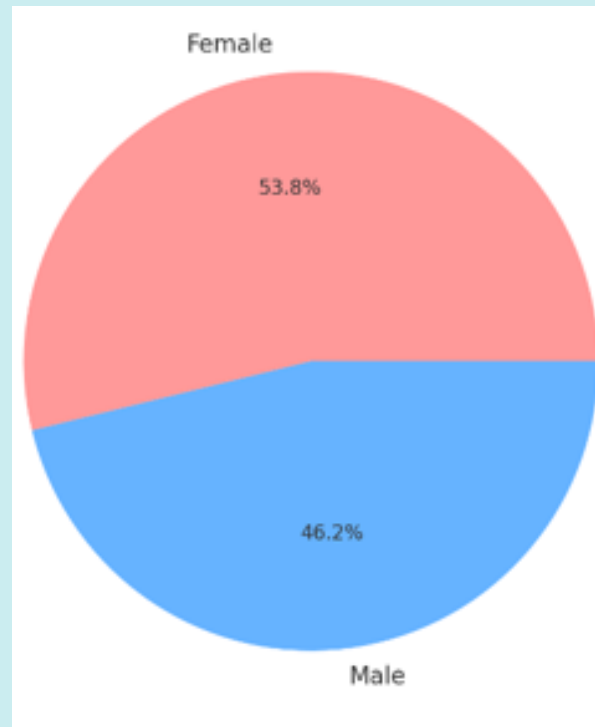
The conducted focus group totals over 50 young individuals aged between 20 and 26, including university students and local organization volunteers. Across all groups, Instagram and TikTok emerged as the most frequently used social media platforms, primarily for entertainment and socializing. However, these platforms, alongside Facebook, X (Twitter), and Threads, were consistently identified as common spaces for encountering hate speech. Participants reported a noticeable increase in hate speech over recent years, often targeting individuals or groups based on gender, race, ethnicity, religion, body image, and disability. Public accounts and anonymous users were frequently linked to the spread of such content, with hate often hiding behind humor, memes, or ambiguous language. Although some participants had personally experienced hate speech, particularly sexism, xenophobia, racism, and religious intolerance, most were more frequently witnesses than direct targets.

The responses to hate speech were largely passive: blocking, muting, ignoring, or occasionally reporting perpetrators. However, reporting mechanisms were widely perceived as ineffective or frustrating due to vague categories, lack of transparency, and automated responses. Very few participants engaged in direct discussions to confront hate speech, often due to a lack of trust in the platforms' accountability or concern for their own emotional safety. Despite widespread skepticism about current moderation efforts, there was near-unanimous support for stricter policies. Participants advocated for stronger penalties such as temporary or permanent bans, identity verification requirements, and even legal consequences like fines. Many emphasized the need for social media platforms to take more responsibility and prioritize user safety over profit. Moreover, participants envisioned diverse and nuanced tools to address hate speech. Suggestions included apps or extensions that automatically detect and block hateful language before it is posted, offer educational resources and emotional support for victims, improve transparency in content moderation processes, allow user feedback or ratings on problematic accounts, offer opt-in features to limit who can comment or engage with users. Some interviewees stressed the importance of building community-based support systems and educational initiatives to address hate at its root. Others expressed cynicism, arguing that hate cannot be eradicated entirely but only managed more effectively. Notably, many used breaks from social media, conversations with friends or family, and offline coping strategies to recover from distressing online experiences.

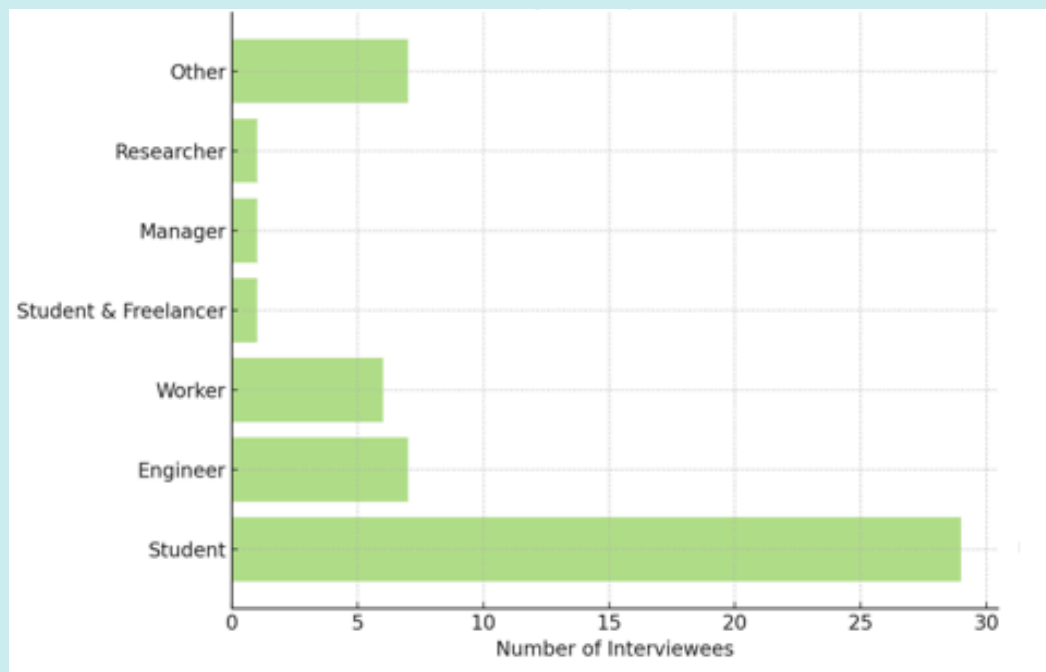
Category	Belgium	Jordan	Tunisia	Italy	Morocco	France
Key Features of Hate Speech	Humor, memes, and sarcasm mask racism and sexism; normalization through satire	Highly politicized; religious intolerance, xenophobia, gender discrimination	Emotionally charged, exclusionary; reflects social inequalities	Racist and anti-migrant rhetoric; intensified during crises	Gender-based attacks, racism, and anti-minority sentiments	Secular vs religious tensions; racialized hate, especially targeting Muslims and immigrants
Common Targets	Muslims, refugees, Black communities, women, LGBTQ+, people with disabilities	Syrian refugees, women, secular voices, activists, minorities	Black Tunisians, migrants, women, LGBTQ+, people with disabilities	Migrants, Roma, Muslims, women, activists	Feminists, Amazigh activists, sub-Saharan migrants, LGBTQ+ people	Muslims, migrants (esp. from North Africa), Black communities, Jews, feminists
Trigger Events	Elections, political debates, global crises	Religious events, legal reforms, political protests	Elections, controversial laws, instability	National political scandals, migration surges	Human rights protests, social reform discussions	Terrorist attacks, political elections, identity debates
Platform Dynamics	Facebook, Twitter, Instagram, Reddit; ineffective moderation; humor shields hate	Twitter, Instagram, TikTok; Arabic-language moderation weak; anonymity emboldens hate	Instagram, TikTok, Discord, Facebook; slow, inconsistent moderation	Facebook, Telegram, TikTok; polarizing narratives spread quickly	Facebook, WhatsApp, TikTok; limited content moderation in Arabic and Darija	Twitter, YouTube, Facebook; reports of biased and inconsistent enforcement
Youth Experiences	Encounter trivialized racism, body shaming, identity-based hate; distrust in platform moderation	90% experience hate; coping through humor, counter-speech, or self-protection	Mental health strain; low trust in platforms; withdrawal common	Youth feel exposed and unsupported; report high emotional impact	High emotional toll; often avoid posting opinions; solidarity groups emerging	Youth politically vocal but often silenced or targeted; Islamophobia and racism normalize during public discourse
Coping & Resistance Strategies	Counter-speech, blocking, creative content, self-care, advocating for better moderation	Humor, peer support, sentiment analysis tools proposed, anonymous reporting	Digital detox, emotional distancing, community support; suggest AI moderation and local language tools	Awareness campaigns, digital literacy education, youth advocacy groups	Education on digital citizenship; safe space forums; youth-driven reporting initiatives	Street protests, online advocacy, fact-checking groups, intersectional activism
Solutions Proposed by Youth	Digital literacy, better reporting tools, real-time AI moderation, transparency	Arabic AI tools, safer reporting systems, peer support features, mental health integration	More transparent systems, inclusive moderation, educational campaigns	Civic education, stronger legal frameworks, platform accountability	Multilingual moderation, culturally aware guidelines, youth co-designing features	Algorithm transparency, anti-racist education, support for online victims, stronger regulations

Interviews data

GENDER DISTRIBUTION

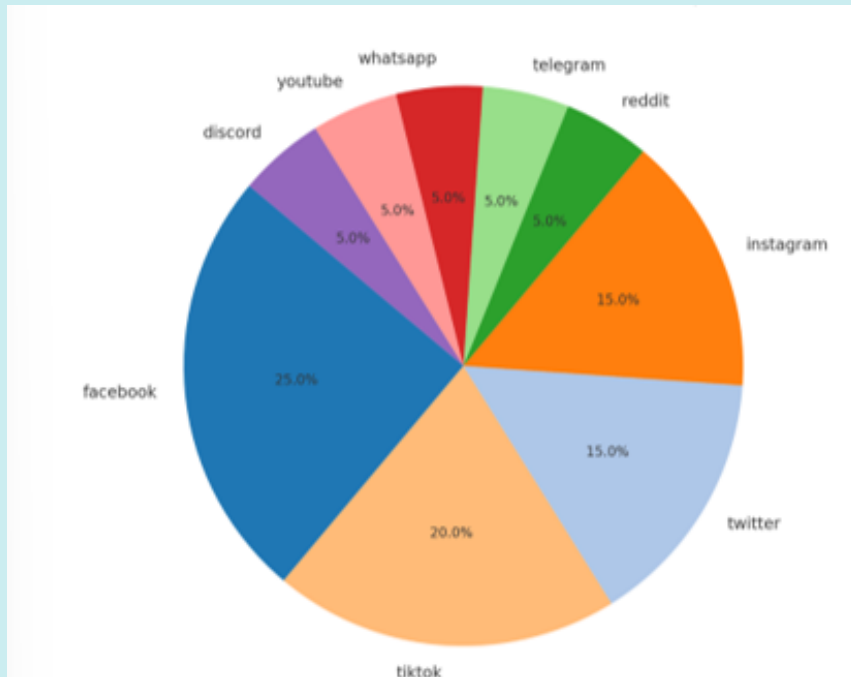


TOP OCCUPATIONS

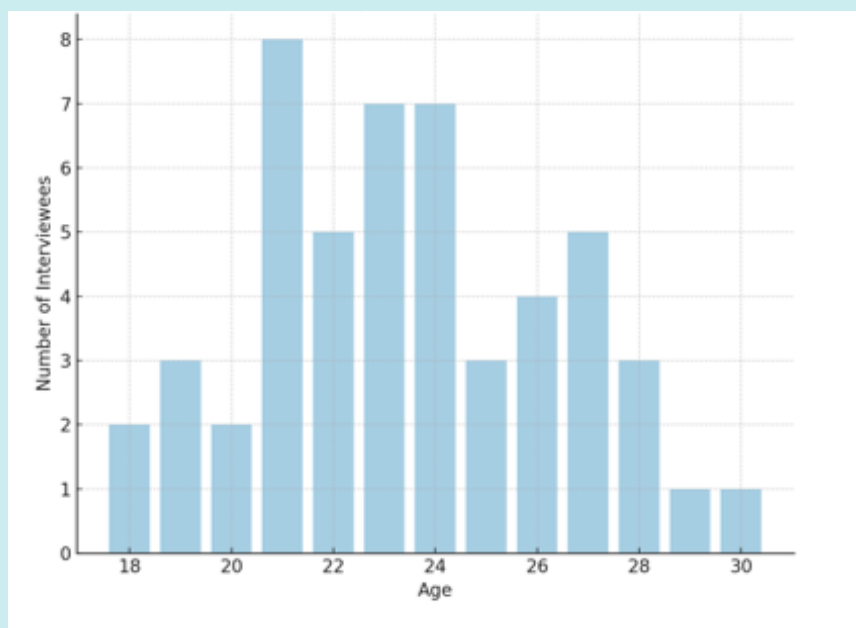


Interviews data

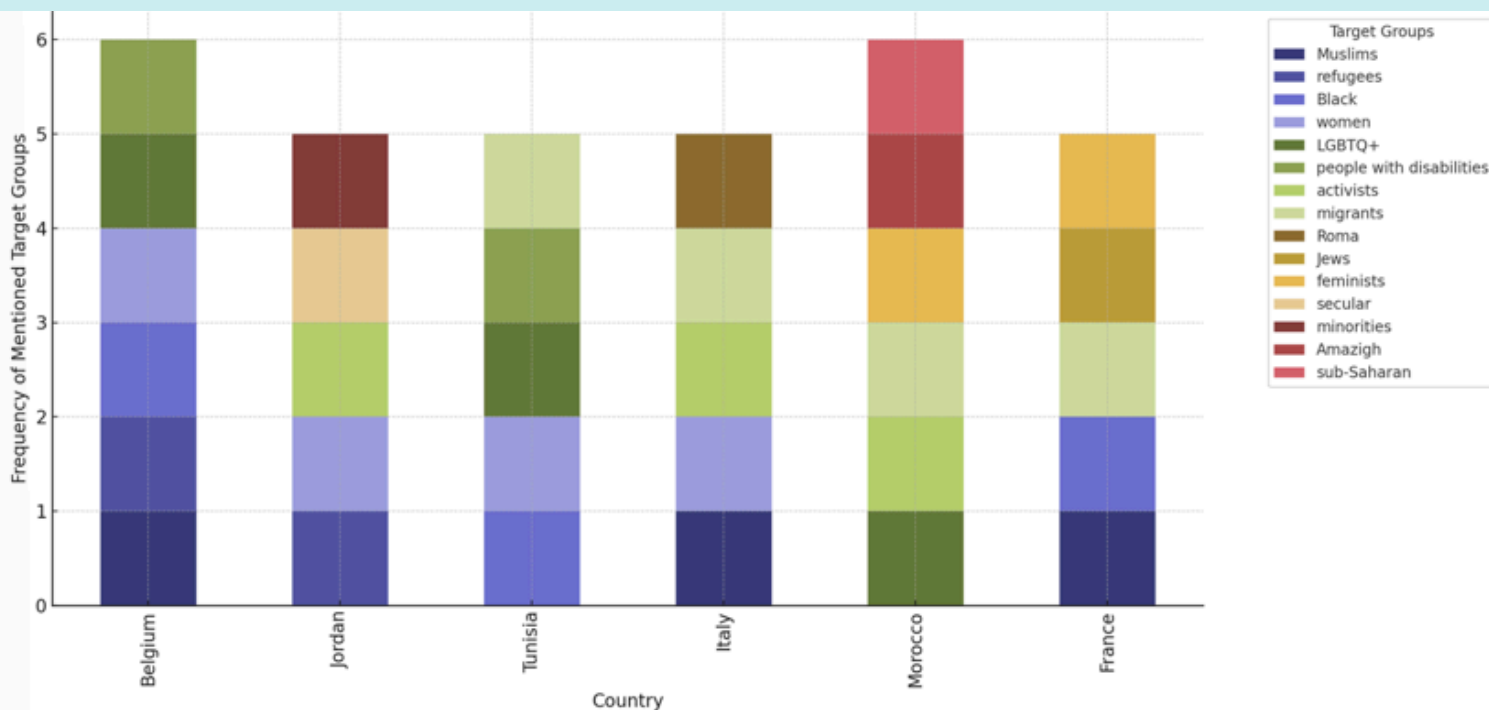
OVERALL DISTRIBUTION OF REFERENCED SOCIAL MEDIA PLATFORMS



AGE DISTRIBUTION



COMMON HATE SPEECH TARGETS BY COUNTRY



3.2 . Discussion

The comparative analysis of Belgium, Jordan, Tunisia, Italy, Morocco, and France reveals a complex, yet deeply interconnected, landscape of online hate speech shaped by historical, cultural, and political contexts. Despite the geographic and societal differences across these countries, several cross-cutting themes emerge, offering valuable insights into how youth experience, respond to, and seek to transform digital spaces. Across all six countries, online hate speech is often embedded in everyday digital practices, normalized through humor, memes, satire, and coded language. In Belgium, for instance, humor is frequently weaponized to mask racist or sexist narratives, blurring the lines between “jokes” and harm. Similarly, in Tunisia and Italy, hate speech is fueled by emotionally charged language and nationalist discourse that justifies exclusion or discrimination. In France, the veil of secularism and national identity is often used to legitimize Islamophobic and anti-immigrant rhetoric, reflecting a broader trend where hate is framed as defense of cultural values.

Common Targets, Local Triggers

While the specific groups targeted differ slightly by context, marginalized identities consistently bear the brunt of online hate: ethnic minorities, migrants, women, LGBTQ+ individuals, and people with disabilities. In Jordan and Tunisia, hate speech often spikes during political events or reform debates, with secular voices, refugees, and gender rights activists facing heightened aggression. In Morocco, online hate is particularly pronounced against feminists, Amazigh activists, and LGBTQ+ voices, often shaped by conservative norms and taboo subjects. Meanwhile, France and Italy see hate intensify around immigration, national crises, or acts of terrorism, revealing how socio-political events act as flashpoints for digital hostility.

Platform Accountability and Trust Deficits

A near-universal frustration with social media platforms emerges in all countries studied. Young people across contexts report moderation systems as ineffective, inconsistent, and opaque. Arabic-speaking users in Jordan, Tunisia, and Morocco face an added layer of exclusion, with AI moderation tools failing to detect or address hate speech in their languages or dialects. The perceived prioritization of engagement and profit over user safety leads to widespread mistrust. Even when users report harmful content, the lack of response, or arbitrary removals, contributes to a sense of helplessness and institutional complicity.

Emotional and Psychological Toll

The cumulative emotional impact of hate speech is profound. Feelings of alienation, anxiety, distress, and fatigue are frequently reported, particularly among young women, activists, and ethnic or religious minorities. In countries like Tunisia, Italy, and Belgium, this emotional toll leads some youth to disengage from digital platforms altogether, viewing them as toxic spaces. Others, such as participants in Jordan and France, express a heightened sense of polarization and division, particularly during politically sensitive periods.

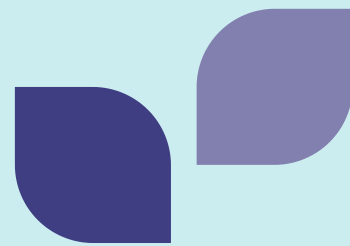
Youth Resilience and Imaginative Resistance

Despite the harms experienced, young people are not passive victims of online hate. Across all six countries, they demonstrate resilience, creativity, and a clear desire for structural change. Youth actively engage in counter-speech, educational initiatives, peer support, and advocacy. They suggest innovative solutions, ranging from AI-powered moderation and anonymous reporting tools to community-led monitoring systems and mental health integration within platforms. Young participants also emphasize the importance of digital literacy and intersectional education in schools as a preventive approach. Particularly noteworthy are calls for moderation tools that are sensitive to local languages, cultures, and political contexts—a rejection of one-size-fits-all moderation models.

Concluding Reflections

The findings from these countries underscore that online hate speech is not merely a technological issue but a reflection of broader societal inequalities and fractures. It flourishes where systems fail to protect the vulnerable, and where discourse becomes a battleground for identity, power, and belonging. Youth across Belgium, Jordan, Tunisia, Italy, Morocco, and France are keenly aware of these dynamics. Their lived experiences offer a sobering yet hopeful roadmap: one that calls for urgent reform in platform governance, investment in inclusive digital tools, and the amplification of youth voices as key agents of change. If digital spaces are to become truly inclusive and safe, they must be co-designed with and for the diverse communities that inhabit them.

4. Conclusions and final recommendations



The comparative insights from Belgium, Jordan, Tunisia, Italy, Morocco, and France reveal that online hate speech is a deeply entrenched phenomenon, rooted as much in offline societal tensions as in the structures of digital platforms themselves. Despite differences in historical and cultural contexts, young people across these diverse settings encounter a digital environment where hate is often normalized, overlooked, and, at times, strategically weaponized. Common patterns, such as the targeting of marginalized groups, the emotional burden of constant exposure to hate, and deep frustrations with platform accountability, point to systemic failures in both societal and technological responses. Yet, amid these challenges, youth demonstrate extraordinary resilience, coping strategies and agency. They are not only resisting hate but actively imagining and building alternatives through counter-speech, education, advocacy, and calls for culturally nuanced, community-driven platform reforms. Ultimately, tackling online hate speech demands more than better algorithms or stricter regulations; it requires a fundamental rethinking of digital governance through an intersectional, participatory lens. Platforms must center the lived realities of young people, particularly those from marginalized backgrounds, and invest in inclusive, context-sensitive solutions. In doing so, digital spaces can evolve from battlegrounds of division into spaces of solidarity, creativity, and transformative change.