EUROPEAN UNION
ELECTION OBSERVATION

GUIDANCE NOTE FOR DEFINING 'HATE SPEECH' IN THE ELECTORAL CONTEXT

DECEMBER 2025

Contents

Introduction and context

Violent language in politics is a growing reality in many countries where EU Election Observation Missions (EU EOMs) are deployed. Cyberspace has increased the harmful effects of some expressions or forms of violent communication[1].

However, not all violent language can be considered as "hate speech". Although there is no international legal definition of hate speech, a number of definitions and approaches to the concept are current under discussion.

Why it is important to define and analyse hate speech, in the electoral context?

→ As the United Nations Secretary General stated, "*addressing hate speech does not mean limiting or prohibiting freedom of speech. It means keep hate speech from escalating into something more dangerous, particularly incitement to discrimination, hostility and violence, which is prohibited under international law*". That is why hate speech is a menace to democratic values, social stability and peace[2].

→ According to the Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (2019)[3], hate speech has a double ambiguity: "*its vagueness and the lack of consensus around its meaning can be abused to enable infringements on a wide range of lawful expression. During electoral process, politicians are increasingly using cyberspace to conduct electoral campaigns, to capture the attention and voting preference of the citizens. Therefore, it is key to observe and analyse violent language in the electoral context*."

→ Online communication amplifies these risks. The EODS Digital Toolkit introduces online impact as an indicator of how far and how strongly content stands out on a platform. Online impact combines reach, engagement and virality, and helps prioritise which items may require closer scrutiny. Information manipulation during elections can be defined as the unjustified and illegitimate use of methods to influence public opinion and voters' choices, thereby reducing citizens' ability to exercise their political rights.

→ It is recognised that no issue is more problematic for those concerned with media freedom than the issue of *hate speech*, as ACE Electoral Knowledge Network[4] acknowledges. The term is generally used to refer to the advocacy of hatred based on national, racial, religious, or other grounds. The question is: "*how far it is proper or acceptable to limit the right to freedom of expression, when the views being expressed support the limitation or infringement of the rights of others".*

[1] Taxonomy of violent communication and the discourse of hate on the internet. Revista de Internet, Derecho y Política (No 22, June 2016, pages 82-107).
[2] Strategic and Action Plan on Hate Speech (2019).
[3] Resolution A/74/486 (2019). Promotion and protection of the right to freedom of opinion and expression.
[4] The ACE Electoral Knowledge Network is the world's largest online community and repository of electoral knowledge. It provides comprehensive information and specialised advice on any aspect of electoral processes.

→ The observation of hate speech becomes more relevant during election processes for two main reasons: i) an election is the moment when a variety of political views should be expressed. To limit expression of some of these views potentially limits not only rights of free speech but also rights of democratic participation; ii) the highly charged atmosphere of an election campaign may be precisely the moment when *inflammatory statements* are likely to have the effect of inciting people to violence - thereby infringing the democratic and free speech rights of others[5].

→ The European Commission Code of Conduct on Countering Illegal Hate Speech Online (June 2016) highlights that digitalisation has amplified citizens' vulnerability to hate speech and disinformation, enhancing the capacity of state and non-state actors to undermine freedom of expression[6]. Disinformation impacts the democratic process and human rights.[7]

All these elements - incitement to violence, violence in an electoral context and digitalisation - show the importance of having a basic guidance note for EU election missions to identify hate speech most consistently. **This Guidance Note provide tools and definitions to better understand what we can and cannot consider hate speech in the electoral context, and to observe and analyse hate speech and** other related violent speech and communication.

Section II provides a set of definitions of hate speech with their references, with the intention of clarifying, facilitating and harmonising the activities of observation and analysis, especially when assessing freedom of expression.

This document is a "living document" and it would need to be updated to reflect emerging trends and the evolution of international standards. The application of the Guidance Note in each mission could offer lessons learned and good practices. Collaboration among political, legal, media, and social media analysts during election missions will help refine the Guidance Note and make them increasingly useful.

This Guidance Note should be read alongside the EODS Toolkit, which introduces additional concepts used during social media monitoring. These include initial assessments of online impact, based on reach, engagement, and virality, as well as an initial evaluation of the potential to cause harm. These tools do not replace the legal and definitional framework outlined above. Instead, they support the prioritisation of cases during monitoring and help analysts determine which incidents require further investigation.

[5] Media and Elections. ACE. Policies on hate speech and defamation.
[6] Code of Conduct on Countering Illegal Hate Speech Online, June 2016. The European Commission launched the Code of Conduct in May 2016 together with four major IT companies (Facebook, Microsoft, Twitter and YouTube)
[7] The impact of disinformation on democratic processes and human rights in the world. European Parliament. Study April 2021.
https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU(2021)653635_EN.pdf

Definitions and references related to hate speech

According to the UN Strategy and Plan of Action on Hate Speech, there is no international legal definition of hate speech, and the characterisation of what is "hateful" is controversial and disputed.

A) United Nations international references

International Convention on the Elimination of All Forms of Racial Discrimination (1965, entry into force 1969)
Article 5 states that States parties condemn all propaganda and all organizations which are based on ideas or theories of superiority of one race or group of persons of one colour or ethnic origin, or which attempt to justify or promote racial hatred and discrimination in any form. States parties should undertake to adopt immediate and positive measures designed to eradicate all incitement to, or acts of, such discrimination and, to this end, with due regard to the principles embodied in the Universal Declaration of Human Rights.

International Covenant on Civil and Political Rights (1966), entry into force 1976) The ICCPR sets key international standards relevant to what is commonly referred to as "hate speech," primarily through Articles 19 and 20. Article 19 protects freedom of expression, while allowing certain restrictions that are provided by law and necessary for the protection of the rights of others or public order. Article 20 establishes mandatory prohibitions. Article 20(2) states: "Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law."
Although the ICCPR does not define "hate speech" as a standalone concept, Article 20(2) creates a binding obligation for States to prohibit advocacy of hatred that amounts to incitement to discrimination, hostility, or violence, thus requiring a balance between freedom of expression and the prevention of serious harm.

Rabat Plan of Action (2017)
Without a specific definition of hate speech, it recommends that a clear distinction be made between (a) expression that constitutes a criminal offence, (b) expression that is not criminally punishable, but may justify a civil suit or administrative sanctions and (c) expression that does not give rise to any of these sanctions but still raises concern in terms of tolerance, civility and respect for the rights of others. The Rabat Plan of Action endorses the Camden Principles on Freedom of Expression and Equality, which sets out the moral and social responsibilities that the media, politicians, religious leaders and civil society each have to combat intolerance.

United Nations Strategy and Plan of Action on Hate Speech (2019)
The term *"hate speech is understood as any kind of communication in speech, writing or behaviour, 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, colour, descent, gender or other identity factor".*

One of the most important aspects of these basic definitions is the listing of the main categories that identify hate speech (nationality, race, religion, etc), while the clause "other identity factor" allows the concept to be further elaborated and updated.

Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (2019)
According to the Special Rapporteur, under international human rights law, the limitation of hate speech seems to demand a reconciliation of two sets of values: democratic society's requirements to allow open debate and individual autonomy and development with the also compelling obligation to prevent attacks on vulnerable communities and ensure the equal and non-discriminatory participation of all individuals in public life. Tech company definitions of hate speech are in general difficult to understand[8].

B) Regional references:

Jurisprudence of the European Court of Human rights. [9]
European Convention on Human Rights (1950, entry into force on 1953)
According to the ECHR, to label hate speech, six parameters must be analysed: the subject matter of the message; who is the sender of the message; the intention of the sender; the target group of the speech; the geographical area where the message is disseminated; and the channel used to disseminate the message.

In other words, the existence of hate messages, even repeated ones, would not *per se* constitute hate speech. It should be noted that the ECtHR also distinguishes between two types of hate speech according to its case law: A. Judgments in which the European Court requires that the hate speech be capable of provoking immediate physical violence against a group. B. Judgments in which the European Court considers hate speech to be that which is capable of promoting immediate physical violence against a group in the long term[10].

American Convention on Human Rights (1969, entry into force on 1978). **Organization of American States.**
Article 13.5 : "All propaganda for war and advocacy of national, racial or religious hatred that constitutes incitement to violence or any other similar unlawful action against any person or group of persons on any grounds including those of race, colour, religion, language or national origin shall be prohibited by law.

[8] In some tech companies such definitions are non-existent, and in others they are vague. Examples: Russian social network VK, the Chinese messaging app WeChat, META. While they use different terms to signal the restriction of content that "promotes" violence or hatred against specific protected groups, they do not clarify how they define promotion, incitement, targeting groups and so forth.
[9] Interesting references for jurisprudence of the European Court of Human Rights freedom of expression and hate speech available at **Factsheet on Hate Speech – European Court of Human Rights (November 2023).**
[10] Hate speech on social media: freedom of expression at a crossroads. Authors: Laura Díez Bueso. Revista catalana de dret públic, Nº. 61, 2020, pp. 50-65 .

Declaration of Principles on Freedom of Expression Organization of American States (OAS) (2000) It was adopted by the Inter-American Commission on Human Rights (IACHR), through its Special Rapporteur for Freedom of Expression who indicates that " *every person has the right to seek, receive and impart information and opinions freely in the terms stipulated in article 13 of the American Convention on Human Rights. All persons shall have an equal opportunity to receive, seek and impart information through any media without discrimination on any ground, including race, colour, religion, sex, language, political or other opinion, national or social origin, property, birth or other status*."

General Policy Recommendation nº 15 on hate speech (European Commission Against Racism and Intolerance (ECRI) (2015)
"*Hate speech is based on the unjustified assumption that a person or a group of persons are superior to others; it incites acts of violence or discrimination, thus undermining respect for minority groups and damaging social cohesion."* In this recommendation, anti-hate speech measures must be well-founded, proportionate, non-discriminatory, and not be misused to curb freedom of expression or assembly nor to suppress criticism of official policies, political opposition and religious beliefs. Hate speech for the purpose of the Recommendation entails the use of one or more particular forms of expression – namely, the advocacy, promotion or incitement of the denigration, hatred or vilification of a person or group of persons, as well any harassment, insult, negative stereotyping, stigmatisation or threat of such person or persons and any justification of all these forms of expression – that is based on a non-exhaustive list of personal characteristics or status that includes "race", colour, language, religion or belief, nationality or national or ethnic origin, as well as descent, age, disability, sex, gender, gender identity and sexual orientation.

The Recommendation specifically excludes from the definition of hate speech any form of expression – such as satire or objectively based news reporting and analysis - that merely offends, hurts or distresses.

African Commission on Human and People's Rights Resolution on the Right to Freedom of Information and Expression on the Internet in Africa (Resolution 362/2016)
In its Resolution, the Commission condemns the use of hate speech on the Internet, such as any form of speech which degrades others, promotes hatred and encourages violence against a group on the basis of criteria including race, colour, religion, national origin, gender, disability or a number of other traits.

Recommendation on combating hate speech of the Committee of Ministers of the Council of Europe (2022)
For the purposes of this Recommendation, hate speech "*is defined as any type of expression that incites, promotes, disseminates or justifies violence, hatred or discrimination against, or denigrates, a person or a group of persons on the basis of their actual or perceived personal characteristics or status such as 'race', colour, language, religion, nationality, national or ethnic origin, age, disability, sex, gender identity and sexual orientation".*

Given that hate speech covers a range of hateful expressions, which differ in their severity, the harm they cause and their impact on members of particular groups in a variety of contexts, Member States should ensure that a properly calibrated set of measures is in place to effectively prevent and combat hate speech. This approach, which needs to be consistent with the European Convention on Human Rights, should distinguish between: 1) hate speech prohibited by criminal law and hate speech which does not reach the level of seriousness required to give rise to criminal liability, but which nevertheless falls within the scope of civil or administrative law; 2) offensive or harmful forms of expression which are not sufficiently serious to be legitimately restricted under the European Convention on Human Rights, but which nevertheless require alternative responses such as: counter-speech and other counter-measures; measures to promote intercultural dialogue and understanding, including through the media and social networking; and relevant education, information-sharing and awareness-raising activities.

Finally, it is important to be aware of the definitions of hate speech provided by social media platforms such as Meta. According to **META Definitions**, their current concept of hate speech is anything that directly attacks people based on what are known as their "protected characteristics"- race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender, gender identity, or serious disability or disease.

For operational monitoring, EU EOMs classify content into three categories: hate speech, defamation and violent communication. This simplification ensures consistent coding and aligns the Guidance Note with the Methodological Checklist.

SECTION I: Overview of Hate Speech Definitions

This section summarises how EU election missions apply the definitions presented when assessing speech during an electoral period. The purpose of this catalogue is not to introduce new terminology, but to clarify how the core concepts operate within the observation methodology. For monitoring purposes, the EU EOM Social Media Analyst (SMA) groups relevant forms of harmful or hostile speech into three operational categories: **Hate Speech, Violent Communication and Defamation.**

These categories translate the broader international framework into a practical approach for assessing online and offline speech during EU Election Observation Missions.

- Hate Speech refers to identity-based attacks, as defined in the international standards cited in this guidance note.
- Violent Communication encompasses hostile or aggressive expressions that are not identity-based.

- Defamation concerns false statements of fact that harm reputation.

These categories are not legal findings. They are analytical tools that support consistent observation and enable the mission to distinguish identity-based harm from broader abusive or misleading communication. As the Toolkit emphasises, this operational simplification enables standardised tagging, more transparent reporting, and comparability across missions.

1. Other related areas: defamation, freedom of expression, synonyms.

Hate speech and defamation

Defamation is often associated with hate speech. The International Press Institute highlights that a growing number of international authorities on freedom of expression have called on governments to abolish or consider abolishing criminal defamation[11]. Generally, defamation is a "false and unprivileged statement of fact that is harmful to someone's reputation, and published "with fault," meaning as a result of negligence or malice."[12] State laws often define defamation in specific ways.

[11] These authorities include the U.N. Human Rights Committee, which is responsible for interpreting the International Covenant on Civil and Political Rights; and the special representatives on freedom of expression of the U.N., the Organization for Security and Co-operation in Europe (OSCE), and the Organization of American States (OAS). While the European Court of Human Rights (ECtHR) has never explicitly ruled out the use of criminal laws in defamation cases, it has criticised their use and suggested that the appropriate space for their use, if any, is narrow. In any case, the ECtHR has joined a very clear international consensus against even the possibility of prison sentences in defamation cases. The Inter-American Court of Human Rights and the African Court on Human and People's Rights have also issued decisions criticising the application of criminal defamation laws.

[12] Media Defence, *What Is Defamation?,* Introductory Modules on Digital Rights and Freedom of Expression Online, Module 5, available at:https://www.mediadefence.org/ereader/publications/introductory-modules-on-digital-rights-and-freedom-of-expression-online/module-5-defamation/what-is-defamation/

Libel is a written defamation; slander is a spoken defamation[13]. The basis for the notion of defamation in international law is article 17 of the International Covenant on Civil and Political Rights (ICCPR), which provides for protection against unlawful attacks on a person's honour and reputation. Article 19.3 of the ICCPR also refers to the rights and reputation of others as a legitimate ground for limitation of the right to freedom of expression[14].

Hate speech and freedom of expression

Freedom of expression protects statements of opinion regardless of their implicit value or truth. Codes of Conduct agreed between political parties in advance of an election campaign can address hate speech[15]. The European Court of Human Rights "has held that truth is an absolute defence to a suit of defamation. That is, if something is true, it cannot be defamatory"[16].

In this matter, the ECHR has an extended jurisprudence of allegations of "freedom of expressions" attempting to cover up a "hate speech". It is useful that the last factsheet published by the ECtHR (November 2023), in which a compilation of this jurisprudence is presented.

Synonyms

It is important not to get lost on synonyms or in closed definitions of similar concepts. Notions such as hate, violence, harassment, insult… have a multitude of synonyms that can interfere with the observation and the analysis of hate speech. That is why all prescriptions must be interpreted in accordance with the national/local context. It would be futile to seek differences between adjectives that are very close to one another, since most of the time these differences reflect personal, intimate, cultural, or contextual interpretations, among other factors.

Inflammatory speech is derogatory speech and becomes hate speech specifically because of its propagation. Derogatory[17] has as a synonym the words insulting, pejorative, malicious, abusive, degrading, hateful, denigratory, slanderous, defamatory, among others. The term "inflammatory" can also be used to denote incendiary, dangerous, or agitational. Harassment can be interpreted as persecution, an offence, or a disturbance, among other things. "Violent" can be defined as brutal, cruel, vicious, coercive, or inflamed.

---

[13] Electronic Frontier Foundation, Online Defamation Law, *Legal Guide for Bloggers*, available at: https://www.eff.org/issues/bloggers/legal/liability/defamation?

[14] Media Defence, *What Is Defamation?,* Introductory Modules on Digital Rights and Freedom of Expression Online, Module 5, available at:https://www.mediadefence.org/ereader/publications/introductory-modules-on-digital-rights-and-freedom-of-expression-online/module-5-defamation/what-is-defamation/

[15] Sometimes, as in South Africa and Cambodia, such a code will have the effect of law.

[16] International Press Institute. International standards.

[17] The **Dangerous Speech Project** preferred to avoid the use of the familiar term "hate speech" as for them, the term is vague, broad, and in practice, everyone defines it differently. The Project defines the so-called dangerous speech as "any form of expression (e.g. speech, text, or images) that can increase the risk that its audience will condone or participate in violence against members of another group".

2. How to identify "hate speech" (Key elements)

Hate speech involves understanding its key elements and characteristics; definitions may vary slightly across cultural, legal, and social contexts. Taking into account all the references, we can consider various aspects for identifying hate speech in the electoral context. These include basic elements such as: what? how? who?, among others.

What?

Dehumanisation or Stereotyping: Hate speech often involves dehumanising language or stereotypes that portray targeted individuals or groups as inferior, dangerous, or unworthy of respect. This can involve using derogatory terms, slurs, or negative generalisations.
The analysis should determine whether the speech targets protected characteristics commonly associated with hate speech, including religion, ethnicity, nationality, race, colour, descent and gender.

How?

Incite Violence: Hate speech is usually intended to harm, intimidate, or incite violence against the targeted individuals or groups. It may express hostility, prejudice, or animosity towards them, contributing to a climate of fear and discrimination. Another element to consider is the **type of action** expressed in the speech, which may include denigration, hatred, vilification, harassment, insults, negative stereotyping or stigmatisation. **Incitement to violence** is a particularly serious and common form of hate speech.

To whom and from whom?

Targeted Identity or Group: Hate speech typically targets individuals or groups based on characteristics such as race, ethnicity, religion, nationality, gender, sexual orientation, disability, or other protected characteristics.

Context and Power Dynamics: Hate speech often occurs within a context of power dynamics, where individuals or groups with more social, economic, or political power use language to marginalise, oppress, or discriminate against those with less power. Analysing from whom it comes from is important; for example, to identify if it is a natural or legal person, public or private institution, media, political party, etc.

Other aspects:

→ Public Expression: Hate speech is frequently disseminated publicly, including through verbal communication, written text, social media posts, online forums, and other media. Its public nature does amplify its harmful effects and contributes to the normalisation of discriminatory attitudes and behaviours.

→ Historical and Societal Context: Hate speech can be shaped by historical and societal factors, including systemic inequalities, social prejudices, and cultural norms that perpetuate discriminatory attitudes and behaviours.

→ Frequency: This aspect concerns the impact of hate speech. For that, it is necessary to check the frequency, analyse if the speech is systematic, done or acting according to a fixed plan or structure, or coordinated and methodical. ***This last element is important to distinguish a "hate message", unique and isolated, from "hate speech", which is permanent and/or consistent over time.***

Remember: Not all violent language can be assessed as hate speech. Isolated hate messages are not hate speech. For example, in order to conclude that "hate speech" is present in an electoral campaign it is necessary to take into account all the elements also those related to the national context.

SECTION II. Observing Hate Speech during EU EOMs

EU Election Observation Missions assess hateful and harmful speech within a broader information environment. Observation, therefore, includes both the content itself and the way it circulates online. The EODS Digital Toolkit sets out three complementary elements that guide this work.

1. Post-level online impact

Online impact provides an initial snapshot of how far and how strongly a piece of content stands out on a platform. As outlined in the Toolkit, this combines:

- **Reach** – how many people are likely to have seen the content.
- **Engagement** – how many people interacted with it.
- **Virality** – how fast it is growing relative to the account's usual performance.

These indicators are always interpreted relative to the platform, the actor and the intended audience. Their purpose is not to determine whether speech is hateful, but to help identify which incidents may require further investigation.

2. Initial assessment of potential to harm

Potential harm is a preliminary judgement about what the content could realistically cause if amplified or repeated. The Toolkit identifies three broad levels: **High**, **medium** and **low** potential to harm.

Factors considered include:

- Whether the content incites violence or hostility;
- Whether it misleads voters about procedures or eligibility;

- Whether it targets vulnerable groups or rights;
- Whether it risks suppressing participation or fuelling tension.

This assessment is revised once the case is understood in its full context.

3. Case-level analysis: narratives, actors and campaigns

While monitoring begins at the post level, final assessments are made at the **case level**. The Toolkit distinguishes:

- **Narratives** – recurring claims or storylines;
- **Actors or networks** – accounts, pages or clusters driving amplification;
- **Campaigns** – coordinated efforts to mobilise, influence or mislead.

Case-level analysis considers:

- Cross-platform spread;
- Breakout into media or elite discourse;
- Potential offline impact;
- Whether the case contributes to discrimination, hostility or violence.

This approach enables the mission to assess not only *what* was said but also *how far it travelled* and *the risks it posed* to electoral integrity, equality, and participation.

Online impact and potential harm do not replace the analytical criteria described in this Guidance Note. They support the observation process by helping to identify which incidents warrant deeper investigation, based on the scale of exposure and the potential risks posed in the electoral environment.

Indicators

The EODS Digital Toolkit provides indicators to support consistent analysis across online platforms and electoral contexts. These indicators help analysts to:

- Prioritise which content requires deeper investigation;
- Identify emerging patterns of harmful or hateful speech;
- Assess the possible impact on voters, institutions or vulnerable groups.

The leading indicators used during observation include:

A. ***Online impact***

Reach, engagement and virality, interpreted relative to the actor, platform and intended audience.

B. ***Potential to harm***

Initial assessment of whether the content could realistically cause disruption, intimidation, discrimination or misinformation-related risks.

C. ***Amplification patterns***

Repetition, coordinated spread, cross-platform movement, sudden spikes and inauthentic behaviour.

D. ***Narrative-level indicators***

Persistence, breadth of adoption, visibility across political communities, resonance with existing tensions.

E. ***Actor and network indicators***

Influence, audience size, engagement profile, cross-platform presence and coordinated behaviour.

These indicators help determine whether specific incidents remain isolated, form part of a broader narrative or constitute a potential manipulation effort.

Methodological Checklist

Step 1 – Identify the Target and Protected Ground

A. Questions

B. Identity reference
- Does the content refer to a person or group using identity attributes such as religion, ethnicity, nationality, race, colour, descent, language, gender, gender identity, sexual orientation, disability or age?
- Is the identity reference explicit (clearly stated)?
- Is the identity reference implicit (coded language, stereotypes or widely recognised dog-whistles)?

B. Identity-based stereotyping
- Is the group portrayed as inferior, dangerous, problematic or unworthy of equal treatment because of who they are?
- Does the message rely on generalisations or collective blame?
- Is identity used to justify exclusion, fear or hostility?

C. Test for classification relevance
- If the identity factor were removed, would the message still carry the same meaning?
- Is the targeted group historically marginalised or vulnerable in the national context?

PROTECTED GROUND
✦ Religion or belief
✦ Ethnicity
✦ Nationality
✦ Race
✦ Colour
✦ Descent
✦ Language
✦ Gender identity
✦ Sexual orientation
✦ Disability
✦ Age
✦ Other identity factor
✦ Unknown

Classification Guidance: If there is no identity element, the content may still be harmful but should be coded as violent communication or defamation rather than hate speech.

Step 2 - Assess the Form and Severity of Expression

Questions

- What type of expression is used?
- Is it slur-based, dehumanising, stigmatising, threatening or calling for exclusion or violence?
- How severe is the content?

A. Severity Levels

🟢 **Level 1 – Hostile or derogatory**

🟡 **Level 2 – Advocacy or normalisation of discrimination**

🔴 **Level 3 – Advocacy or celebration of violence**

Tags Completed

Type of Expression Tags

TYPE OF EXPRESSION
Slur
Insult
Dehumanisation
Negative stereotype
Denigration
Vilification
Harassment
Threat
Stigmatisation
Call to exclusion
Call to discrimination
Call to violence
Celebration of violence
Manipulated or deceptive content with hateful framing
Other / Unknown

Classification Guidance: Level 3 requires urgent escalation. Levels 1 and 2 still constitute hate speech when linked to protected grounds.

Step 3 - Evaluate Intent, Context and Power Dynamics

Actor-focused assessment

Questions

- Who is speaking and what is their influence?
- Who is targeted, and are they vulnerable or historically marginalised?
- Is the message likely to intimidate or incite discrimination or violence?
- Is the content linked to a tense political moment?

Tags Completed

- Actor category
- Contextual notes

Classification Guidance: A message by a senior political figure targeting a vulnerable group carries more weight than a low-influence citizen post. Power imbalance increases harm and potential for incitement.

Step 4 - Public Nature and Dissemination Pattern

Questions

- Is the content posted publicly or inside a group that reaches many users?
- How widely did it spread?
- Was amplification organic or coordinated?
- Is this a single incident or part of a repeated pattern?

A. ***Platform Tags***

| Facebook | Instagram | TikTok | X | YouTube | Telegram | WhatsApp public groups | Online news sites | Forums | Local platforms | Unknown |

B. ***Format Tags***

| Text | Image | Video | Meme | Audio | Livestream | Mixed format | Unknown |

C. ***Frequency and Pattern Tags***

Pattern	Description
Isolated single incident	One-off occurrence
Repeated incidents by the same actor	Same actor posting multiple occurrences
Repeated incidents by multiple actors	Wider pattern of behaviour
Spontaneous viral spread	Unplanned but rapid reach
Coordinated spread	Signs of organisation or planning
High intensity burst	Short-term spike in activity
Sustained multi-week pattern	Long-term repetition
Unknown	

Classification Guidance: Patterns matter. An isolated remark is documented, but large-scale repetition may indicate a hate-based influence campaign.

Step 5 - Links with Information Manipulation Activities

Use this section to identify overlaps with manipulation tactics.

Questions

- Does the hateful content also contain misleading or deceptive claims?
- Does it mirror misinformation narratives observed elsewhere?
- Is it being used to suppress participation or influence voter choice?

Tags Completed

- Tactic observed
- Contextual notes
- Estimated volume

A. *Tactic Observed (DISARM-Inspired)*

Identity-based framing	Fear framing	Demeaning humour
Scapegoating	Emotional agitation	Forced association
Coordinated repetition	Manipulated media	Misleading reframing
Flooding or spamming	Artificial amplification	Hashtag manipulation
Cross-platform seeding	Other	Unknown

Classification Guidance: Hate-based narratives are often used alongside disinformation to intensify impact.

Step 6 - Classification of Content

Choose one category:

- Hate speech
- Violent communication
- Defamation
- Harmful but non-hate content
- Other

Classification Guidance:

- Only identity-based attacks qualify as **hate speech**.
- **Defamation** concerns false factual allegations about reputation.
- **Violent communication** includes hostile but identity-neutral language.

Step 7 - Document Clearly and Consistently

Every recorded incident must include:

- ☐ Platform, link and date
- ☐ Protected ground
- ☐ Type of expression
- ☐ Severity level
- ☐ Actor category
- ☐ Tactic observed
- ☐ Frequency and pattern
- ☐ Estimated volume
- ☐ Contextual notes
- ☐ Short description or quote
- ☐ Screenshot where possible

Full Tagging Structure (Summary Table)

Tag Category	Purpose
Protected ground	Identifies identity reference
Type of expression	Determines form of hostility
Severity level	Indicates escalation risk
Actor category	Contextualises influence
Platform and format	Locates environment
Tactic observed	Flags manipulation dynamics
Frequency and pattern	Identifies repetition
Estimated volume	Gauges scale
Contextual notes	Adds nuance

Glossary

English

Hate Speech: any kind of communication in speech, writing or behaviour, 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, colour, descent, gender or other identity factor. **United Nations Strategy and Plan of Action on Hate Speech** (2019).

According to the regional and national context in which EU mission will be observing, other references will need to be taken into consideration (please, refer to the Chapter 2).

Violent communication: is communication that limits liberty, denies recognition of needs, diminishes the worth of a person, and/or blocks compassionate. Violent communication is often the result of using manipulative or coercive language that induces fear, guilt, shame, praise, blame, duty, obligation, punishment, and/or reward. Common ways that violent communication occurs are through (among others): moralistic judgments and evaluations of others; denial of responsibility for our own feelings, thoughts, and actions.

Defamation is the public communication of a false statement of fact that harms a person's reputation, made with fault, meaning either through negligence or intentional wrongdoing (malice). The precise definition and legal standards for defamation vary under national laws.

Français

Discours de haine: tout type de communication, qu'il s'agisse d'expression orale ou écrite ou de comportement, constituant une atteinte ou utilisant un langage péjoratif ou discriminatoire à l'égard d'une personne ou d'un groupe en raison de leur identité, en d'autres termes, de l'appartenance religieuse, de l'origine ethnique, de la nationalité, de la race, de la couleur de peau, de l'ascendance, du genre ou d'autres facteurs constitutifs de l'identité

Langage violent : est une communication qui limite la liberté, refuse de reconnaître les besoins, diminue la valeur d'une personne et/ou bloque la compassion. La communication violente est souvent le résultat de l'utilisation d'un langage manipulateur ou coercitif qui induit la peur, la culpabilité, la honte, l'éloge, le blâme, le devoir, l'obligation, la punition et/ou la récompense. Les formes les plus courantes de communication violente sont (entre autres) : les jugements et évaluations moralisateurs des autres ; le déni de la responsabilité de nos propres sentiments, pensées et actions.

Diffamation : La diffamation est une fausse déclaration de fait qui porte atteinte à la réputation d'une personne et qui est publiée « avec faute », c'est-à-dire à la suite d'une négligence ou d'une malveillance.

Español

Discurso de odio: es cualquier forma de comunicación de palabra, por escrito o a través del comportamiento, que sea un ataque o utilice lenguaje peyorativo o discriminatorio en relación con una persona o un grupo sobre la base de quiénes son o, en otras palabras, en razón de su religión, origen étnico, nacionalidad, raza, color, ascendencia, género u otro factor de identidad. Estrategia y Plan de Acción de las Naciones Unidas contra el discurso de odio (2019).

Lenguaje violento: es la comunicación que limita la libertad, niega el reconocimiento de las necesidades, disminuye el valor de una persona y/o bloquea la compasión. La comunicación violenta suele ser el resultado del uso de un lenguaje manipulador o coercitivo que induce al miedo, la culpa, la vergüenza, el elogio, la culpa, el deber, la obligación, el castigo y/o la recompensa. Las formas más comunes de comunicación violenta son (entre otras): juicios moralistas y evaluaciones de los demás; negación de la responsabilidad de nuestros propios sentimientos, pensamientos y acciones.

Difamación: es una declaración de hechos falsa y no privilegiada que perjudica la reputación de alguien y que se publica "con culpa", es decir, como resultado de negligencia o malicia. Las leyes estatales suelen definir la difamación de formas específicas.