Co-funded by the European Union

Co-funded and implemented by the Council of Europe

Toolkit for analysing a case of hate speech

Advanced guide

The main problem in analysing a case of hate speech is that there is not a universal methodology to do so. While the European Court of Human Rights does offer clues as to how the severity of a case of hate speech can be determined, these indications are determined indirectly from its case law and, therefore, the methodologies used by national institutions with attributions in sanctioning hate speech vary to a rather high extent.

The Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility of violence, adopted on 5 October 2012, does recommend that there should be made a clear distinction between hate speech that is criminally punishable, hate speech that is not criminally punishable but would justify civil or administrative sanctions and hate speech that just “raises concerns in terms of tolerance, civility and respect for the rights of others”[1]. In order to make the distinction between the criminally punishable hate speech and the other two type of hate speech, the Rabat Plan of Action proposes a six-part threshold test, that takes into account:

1. the Context of the speech
2. the Speaker
3. the Intent
4. the Content and the form of the speech
5. the Extent of the speech and
6. the Likelihood of the speech to produce immediate actions against its targets.

and offers general recommendation as to what needs to be taken into account for each of these six criteria.

The methodology we are proposing starts from the criteria and recommendations expressed in the Rabat Plan of Action and aims at making them easier to operationalize by adding more sub-criteria and offering more concrete advice as to how these sub-criteria can be evaluated.

[1] Rabat Plan of Action Article 20

This project is funded by the European Union’s Rights, Equality and Citizenship Programme (2014-2020) and the Council of Europe.
This document represents the views of the author only and is their sole responsibility. The European Commission and Council of Europe do not accept any responsibility for the use that may be made of the information it contains.

1. Context

Analysing the context means understanding the social and cultural landscape in which the hate speech subjected to the analysis operates. The determination that needs to be done has to do with how vulnerable the target of the hate speech is from a social, cultural or political perspective. While hate speech can target any social group, it is obvious that members of majority population, with easier access to political rights, education etc. are less vulnerable than potentially marginalised groups who have been subjected to a long history of negative stereotyping, lack of access to all kinds of services and weak political self-determination

The main criteria that need to be taken into account are:

a) Determining whether the group targeted by the expression is a potentially vulnerable group. This is a binary assessment (Yes / No) that can be done by looking into whether the group represents a minority from an ethnical / racial / religious / sexual / gender orientation / social status / other criteria view and lack of a position of equal power. Belonging to a minority that holds a position of power (e.g. large business owners, who are minority on grounds of social status) should yield a “No” answer on this criterion. In case of an overlap of group identities (e.g. the target of the expression is both a large business owner and a member of the Roma community), the criterion should be applied to the group targeted by the expression (it should yield “Yes” if the person was attacked on grounds of being Roma and “No” if the person was attacked on grounds of being a large business owner).

b) Type of acts of violence / discrimination carried out in recent years against the group targeted by the expression. The answer options we are proposing, in order of severity, are the following: “Verbal violence”, “Psychological violence”, “Generalized discrimination by fellow citizens”, “Institutionalized discrimination” “Property destruction”, “Generalized and institutionalized restrictions of human or civil rights”, “physical violence”, “murder motivated by hatred”. Choosing the answer should take into consideration the most severe situations in which members of the group targeted by hate speech have found themselves in recent years and which cannot be considered an isolated case.

c) Extent of negative stereotypes towards the group targeted by the expression. We are proposing a three-level approach, with “Some extent”, “Moderate extent” and “High extent” as the answer options. Little extent means that there are just a few people who hold negative stereotypes against the group, while, at the other end of the spectrum, “High extent” means that negative stereotypes are generalized in the society.

d) Connection of the hate message with the negative stereotypes against the group targeted by the expression. The answer options we are proposing, in order of severity, are the following: “No connection”, “allusions towards negative stereotypes”, “affirmation and / or consolidation of negative stereotypes”.

e) Political representation of the group targeted by the expression. The answer options we are proposing, in order of severity, are the following: “Consolidated political representation”, “In-group political representation”, “Limited political representation”, “Lack of political representation”. “Lack of political representation” should be chosen when there are no well-known elected officials who are self-assumed members of the group targeted by the expression. In-group political representation should be chosen when the only well-known elected officials are exclusively members of a party that was formed with the main goal of representing the group targeted by the expression. “Consolidated political representation” is to be chosen when there are multiple well-known elected officials belonging to the group targeted by the expression and these elected officials are members of political different political parties with different ideologies

f) Extent of movements supporting the group targeted by the expression. The answer options we are proposing, in order of severity, are the following: “Generalized support”, “Moderate support” and “Lack of support”. Lack of support is to be chosen when there are few to none local or national stakeholders (NGOs, academic institutions, influencers, regular citizens etc.) who are regularly and publicly supporting the rights of the group targeted by the expression. At the other end of the spectrum, “Generalized support” is to be chosen when public support is shown regularly and by as many stakeholders as possible.

2. Speaker

This analysis serves the goal of clarifying how likely it is that the speakers’ hate message will be positively received and believed by the audience. Therefore, the analysis looks at clues into the influence the speaker has on the audience to which the message has been presented. It is also important to look into how much the speaker has abandoned his political / social / moral obligations when engaging in hate speech (engaging in hate speech is more serious for public servants, who must not discriminate among citizens, than it is for politicians, who are supposed to act as the voice of their constituency and which might hold rather radical views towards some social groups).

The main criteria to be considered are:

a) Status of the speaker. The answer options we are proposing, in order of severity, are the following: “Regular citizen”, “Political figure”, “Public figure or influencer ”, “Educator”, “Public servant”. They are to be evaluated based on the general perception of the audience of the expression regarding the social status of the speaker. “Regular citizen” is to be chosen when the person engaging in hate speech has no particular social status that would place her or him above the audience from a power relation perspective. “Political figure” is to be chosen for politicians or for people strongly associated with social movements, even when these movements are not organized as political parties (e.g. union leaders, NGO representatives etc.). “Public figure or influencer” should be chosen when the speaker is a well-known figure who does not engage (primarily) in political work. Examples of public figures or influencers would be actors, vloggers, journalists, artists. “Educator” is to be chosen for speakers that are teachers, trainers, university professors etc. “Public servant” is to be chosen when the person engaging in hate speech is supposed to serve any member of the society without discriminating against any of them.

b) Capacity in which the speaker made the statement. This criterion adds a new layer to the previous one. Most of the time, the status of the speaker (as seen by the audience of the message) is the same as the capacity in which they deliver the expression. An example would be a politician delivering a speech in a parliament. However, sometimes, the capacity differs from the status, such as when a politician's private conversation is leaked in the public space, or when a vlogger who is also an educator engages in hate speech during class, rather than on their Tik Tok channel. The answer options we are proposing, in order of severity, are the same as for the status of the speaker: "Regular citizen", "Political figure", "Public figure or influencer", "Educator", "Public servant".

c) Credibility of the speaker among the intended audience of the hate message. The answer options we are proposing, in order of severity are the following: "Little to no credibility", "Limited credibility", "Moderate credibility", "High credibility". This assessment can be sometimes hard to make, especially if the target audience of the hate message is not familiar to the evaluator. However, generally, the more similar the values and beliefs of the target audience are to those assumed or associated to the speaker, the higher the credibility the speaker will likely have. Assessing the values and beliefs of the target audience can be done by estimating who the audience is composed of and relying on the previous experience or the expert knowledge of the person making the evaluation.

d) Credibility of the speaker. Similar to the previous criterion, the credibility of the speaker in the eyes of the audience exposed to the expression is also important to evaluate. However, this is required only in those cases in which the expression has reached audiences beyond the ones initially intended by the speaker. The answer options we are proposing, in order of severity are the following: "Little to no credibility", "Limited credibility", "Moderate credibility", "High credibility" and "The expression did not reach any audiences other than the ones intended".

e) Influence of the speaker on the group targeted by the expression. The answer options we are proposing, in order of severity are the following: "Little to no influence", "Limited influence", "Moderate influence", "High influence". In order to assess this, the evaluator should look into how much damage the actions of the speaker acting in accordance to her / his status can cause to the group targeted by the expression. At one end of the spectrum you would have a regular citizen engaging in hate speech against a group whose members he is likely to never meet, while at the other end you would have a public servant whose daily work involves protecting the human rights of people against whom (s)he is speaking.

3. Assumed Intent

Determining the intent of the speaker can provide extremely valuable information in determining the intensity of the action that needs to be taken against the speaker or to compensate for the expression. While intent can be extremely hard to determine, past actions of the speaker, the way the speaker has selected the audience of the message and the way s(he) reacted after the speech are elements that can be rather easily determined and which offer valuable clues. Also, the messages hidden between the lines can also shed light as to the objectives of the speaker, even though they are harder to determine and doing it relies strongly on the experience of the evaluator

a) Past actions of the speaker with regards to the group targeted by the expression. Looking into the past actions of the speaker towards the group targeted by the expression can reveal whether the speaker holds negative feelings towards the group. If the speaker of a negative expression towards a group has in the past fought for the rights of members of that group and has never done anything detrimental to their interest, then it is highly unlikely that the negative expression was disseminated with bad intentions. The opposite is true for somebody who has always engaged in negative actions against the group they are speaking against. The answer options we are proposing, in order of severity are the following: “Positive actions”, “Mixed actions / no actions”, “Negative actions”. When choosing the answer option, we recommend, when possible, to consider more recent actions of the speaker. In other words, if a speaker used to engage in positive actions towards the group targeted by the expression, but in recent years her / his behavior changed and now engages almost exclusively in negative actions, that this option should be chosen instead of the “Mixed actions / no actions”.
b) Reaction of the speaker after promoting the hate message. The way speakers react after disseminating a hateful narrative can provide clues as to the speaker’s actual intentions. Showing true remorse can hint towards the speaker not actually meaning any harm from the use of the expression, while continuing incitement can consolidate the idea that the hate message was premeditated. The answer options we are proposing, in order of severity are the following: “Apologies offered”, “No reaction”, “Continued incitement”. Here also it is important to read between the lines and try to determine if the apologies offered are sincere, or just a way for the speaker to escape potential sanctions.
c) Probable objectives of the speaker. The answer options we are proposing, in order of severity are the following: “Voicing the concerns of the speaker’s supporters / Academic debate /Promoting or expressing the speaker’s religious believes”, “Improving own image among the target audience of the message”, “Discrediting the group targeted by the expression”, “Limiting the rights of the group targeted by the expression”, “Call to violent action”. If the expression follows multiple objectives, the most severe one should be considered.
d) Intended audience of the hate message. While some hate messages reach a larger audience than the one initially intended, understanding who the speaker wanted to address through their message is key to evaluating their intentions. The reason for this is that different audiences tend to react differently to the specific messages they are being presented. While something could sound sarcastic to a group of people, others might take the thing for granted and act upon it. The answer options we are proposing, in order of severity, are: “audience not likely to have negative feelings towards the targets of the expression”, “audience likely to have negative feelings towards the targets of the expression”, “audience having strong negative feelings towards the targets of the expression”.

4. Content and form

Analysing the content and form of the hate may involve certain critical discourse analysis skills and is not easily quantifiable. The experience of the evaluator is key in determining this part of the analysis. The parameters we are proposing are the following

a) Degree to which the expression is provocative or aggressiveness of the message. The answer options we are proposing, in order of severity are the following: “Low degree of violence”, “Moderate degree of violence”, “High degree of violence”. When analysing the expression, attention must be paid as to whether and to what extent, it contained charged words or phrases that are known to elicit negative reactions in the audience towards the group targeted by the hate message.
b) Form taken by the expression. Some forms of expression benefit from a higher degree of protection than others. It is therefore important to make the distinction between protected and unprotected forms of expression. The most common forms of protected forms of expression are artistic expressions, religious expressions, academic discourse and research and public interest discourse (understood as a critical approach to issues of high public interest). It is important that the evaluator makes the distinction between the speaker actually engaging in forms of expression that are protected and them disguising their speech as one of these protected forms (e.g. racism disguised as academic discourse by citing obscure, obsolete theories generally considered by the academic community as untrustworthy or previously proven wrong or aggressive homophobic speech backed by references to religious texts disguised as religious expressions).
c) How direct was the message? The expression that is being analyzed can be openly hateful, or it can try to just suggest the hateful message by using metaphors or other figures of styles. Openly hateful messages containing calls for action tend to be more easily understood as such by the audiences. Therefore, they also tend to be more severe than the hidden ones. The answer options we are proposing are “Direct” and “Indirect”.
d) Degree to which the message can be considered a call to action. This criterion is to be analysed together with the directness of the message. Some hateful messages tend to convey nothing more than the opinion of the speaker regarding the targets of the expression, while others encourage people to act against those targets either directly or indirectly through varied discursive techniques such as suggesting that the harm brought by the targets is imminent or supported by the elites. The answer options we are proposing are “No call to action”, “Could motivate some people to take action”, “Mentions / suggests actions to be taken against the targets of the expression”.
e) Correlation with other dominant hate narratives. Messages that piggy-back on dominant hateful narratives tend to be more easily accepted by the audiences already favourable to the hateful narratives and so, they can be more harmful. Expressions that aim at creating new hateful narratives are harmful too, but, unless they are being disseminated in a concentrated manner (as evidenced by the

analysis done under chapter 5. Extent of the speech act), they are less likely to be accepted by the audiences. The answer options we are proposing are the following: “No correlation with dominant hate narratives”, “Some correlation with dominant hate narratives”, “Expression of a dominant hate narrative”.

f) Legal status of hate message. Some countries have clear provisions on what types of hate speech are punishable under criminal law. To see a collection of national law provisions on this topic, you can access the country-specific information provided by the members of the International Network against Cyber Hate at this link.

5. Extent of the speech act

Analysing an expression that could constitute hate speech should look beyond what was said, by whom, about who or in which context and also consider the magnitude of the dissemination efforts, or the extent of the hate speech act. This means concentrating on the medium in which the speech has been disseminated, the frequency and the quantity of the material being disseminated, and the extent to which the audience was reached (a measure of the efficiency of the dissemination efforts)

a) Nature of the expression. This is the most basic level of analysis and means identifying if the message was expressed in a public or a private context. Expressions disseminated in private contexts, while potentially revealing of a person’s true views towards the targets of the message, does not aim at producing results for the speaker and is, in fact, protected by the right to one’s privacy. A private context can be considered any setting such as a private party or event, a family setting, or a closed group form of communications, such as a mailing list (where people sharing the same interest have asked to be included) or a closed social network group. An expression initially made in a private context which is then leaked to the public by the speaker or with the speaker’s agreement, should be considered as having taken place in a public context.

b) Means of dissemination. Analysis based on this criterion should look at the channels through which the message has been disseminated and evaluate their potential to reach either large audiences or the audience intended by the speaker. The answer options we are proposing are “Likely inefficient at reaching the intended audience”, “Likely moderately efficient at reaching the intended audience”, “Likely efficient at reaching the intended audience”. As a general rule, written news outlets (printed newspapers, local websites) should be considered less efficient in reaching intended audiences than national radio or television, at least in regard to more senior audiences. On the other hand, social media and new media can be considered more efficient than traditional media when it comes to younger audiences. Public discourses which are then not further distributed by the use of other media should be considered less efficient, as long as they were given in front of a general audience, but can also be considered “likely efficient” when delivered in front of an audience made up of supporters of the speaker or his ideology (e.g. an antisemitic discourse at a far-right rally).

c) Frequency of the dissemination of the hate message. The evaluator should look into how many times / how often the speaker has repeated the hate message, either word by word or by rewording it. The answer options we are proposing, in the order of their severity, are: “Single time dissemination”, “Moderate frequency of dissemination”, “High frequency of dissemination”. The answer options “moderate frequency” and “high frequency” should be decided in the particular communicational context in which the expression is made. Clues as to how much the speaker “pushes” the expression can be found by looking, for example, as to how different are the contexts in which (s)he is bringing the subject up, or how natural / unnatural the subject is being brought up by the speaker.

d) Quantity of disseminated materials. This criterion is easier (or even possible) to analyse for printed materials that are being disseminated, such as flyers, brochures, books etc. The answer options we are proposing, in the order of their severity, are: “Few disseminated materials”, “Moderate number of disseminated material”, “High number of disseminated materials”.

e) Accessibility of hate message. Evaluating the severity of a hate message should also be done by looking at how easy the process of accessing the information has been done by the speaker. There is one thing to have a hateful post on the timeline of the Facebook page managed by the speaker and a whole different thing to have it pinned right at the top of it. The answer options we are proposing are “Low accessibility”, “Medium accessibility” and “High accessibility”.

f) Extent of the reached audience. This criterion is very hard to estimate but offers valuable input in determining the severity of hate speech. It should be analysed keeping in mind the intended audience of the expression, but the general audience should also be taken into account for cases in which a “specialised” message made its way to the general public. Strategies to evaluate the extent of the reached audience can involve “guesstimating” it from the engagements a social media post had from its audience or looking into how many (media) platforms have shared the message and corroborating this information with their usual audience profile. The answer options we are proposing are “Low extent”, “Medium low” and “High extent”.

6. Likelihood of generating violent / discriminatory events

The likelihood of the speech act generating a situation which represents a clear and immediate danger to the social group targeted by the expression is probably the hardest and most important test that an evaluator must perform in order to positively qualify an expression of hate speech as being sufficiently extreme to require a criminal investigation of censorship from state institutions. The aim here is to establish a clear cause – effect relationship between the expression and the potential of the audience to act against the targets of the expression.

a) Effects produced by the hate message. The answer options we are proposing are the following: “No effects produced”, “Audience engaged in verbal violent conduct”, “Audience engaged in violent / discriminatory actions”. In order for an expression to be evaluated as having engaged the audience in verbal violent conduct, a significant number of its audience must engage in hate speech using themes, expressions or ideas from the original expression. However, for an expression to be evaluated as having engaged the audiences in violent / discriminatory actions, it is sufficient that just one member of the audience has engaged in such actions, but a clear connection between the expression and the actions must be established (such as the audience member confessing about having the expression as an inspiration source for his actions).

b) Does the audience have the means to act on the incitement? To determine this, the evaluator must have a good understanding of the audience exposed to the expression. Inciting people to act against a group with whom they are extremely less likely to have contact due, for example, to the geographical distances between the groups is far less serious than inciting a group who has power over another group to act against it (for example, inciting teachers to discriminate against children belonging to a certain social group)

c) Probability of the audience acting on the hate message. While the evaluation based on the previous criterion relies on knowing the socio-demographic characteristics and the dynamic of relations between the audience of an expression and its targets, the evaluation based on this criterion is even harder because the inner resorts and motivations of the intended audience must be intuited. The way the audience has reacted in the past to similar messages could be of use here. The answer options we are proposing are “Low likelihood of action taking place”, “Medium likelihood of action taking place”, “High likelihood of action taking place”.