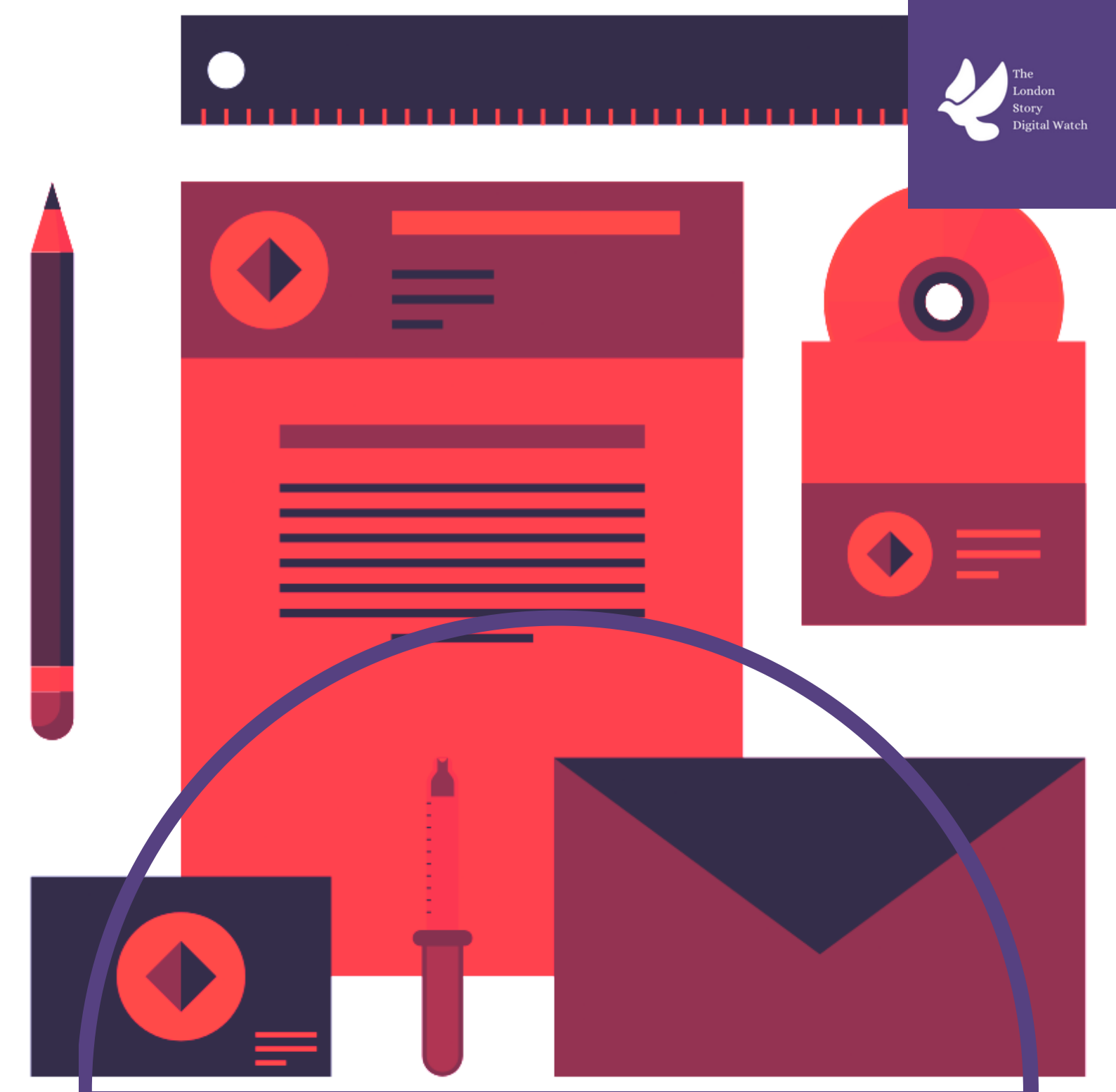


A CIVIL SOCIETY GUIDE TO DOCUMENTING ONLINE HATE SPEECH

STICHTING THE LONDON STORY 2021

Cite as: Alena Kahle and Ritumbra Manuvie (December 2021). *A Civil Society Guide to Documenting Online Hate Speech*. The London Story.

This guide has been prepared by members of Foundation The London Story (TLS), a registered Human Rights Defender Organization in the Netherlands. TLS's overall mandate is to document and inform about human rights violations, and to advocate against hate speech and for Big Tech accountability.

For more information, you can reach us out at **info@thelondonstory.org**

PURPOSE OF GUIDE

Hate speech flourishes on social media. In this guide, we present the methodology that we use for our documentation of hate speech and violent incitement on Facebook and YouTube. Through this guide, we give you the practical tools to create and defend virtual spaces as active citizens.

CONTENT WARNING: We will be using real examples from India of hate speech against religious minorities in this guide for educational purposes.

WHAT DOES THIS GUIDE COVER?

HATE SPEECH

The right to freedom of speech is the most basic human right and the bedrock of democratic societies. Yet worldwide, we are also witnessing a disturbing groundswell of xenophobia, racism and intolerance often promoted through social media platforms.

Platforms like Facebook, Youtube and Twitter are used to spread sensational, misinforming claims and outright bigotry. Such weaponisation of social media has contorted our public discourse with incendiary rhetoric that stigmatizes and dehumanizes minorities, migrants, and women.

Yet, to counter the global phenomenon of hate speech on social media there is no international legal definition of hate speech, and what is considered hateful may vary across nations. For example, the United Nations Strategic Plan of Action on Hate Speech defines hate speech 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."

Similarly, the European Union defines hate speech as "*the public incitement to violence or hatred on the basis of certain characteristics, including race, colour, religion, descent and national or ethnic origin*."

Conversely, in the United States - which houses most of the social media companies - the First Amendment right to freedom of speech can only be infringed upon if it directly incites imminent criminal activity or consists of specific threats of violence targeted against a person or group.

In 2016, the European Commission agreed with Facebook, Microsoft, Twitter and YouTube on a Code of Conduct, and since then, Instagram, Snapchat, Dailymotion, Jeuxvideo.com, TikTok and LinkedIn have followed suit. As part of this Code of Conduct, the companies must have rules and community standards that prohibit hate speech and put systems and teams to review content within 24 hours that is reported to violate the self-regulation standards that platforms adopt.

We rely on the social media platform's definitions to determine when the content is violative of their content moderation policy. To flag content, we pay specific attention to the definition of hate speech, violence, coordinated inauthentic behaviour, and false news. By relying on platform self-regulation, we treat platforms as private entities that have established specific standards for the use of their services. By flagging content against we are essentially holding platforms responsible for what they have promised.

In this guide, we will be relying on the definition of hate speech as adopted by Facebook (an entity of Meta Inc.) and Youtube (an entity of Alphabet Inc.).

DIGITAL ETHNOGRAPHY

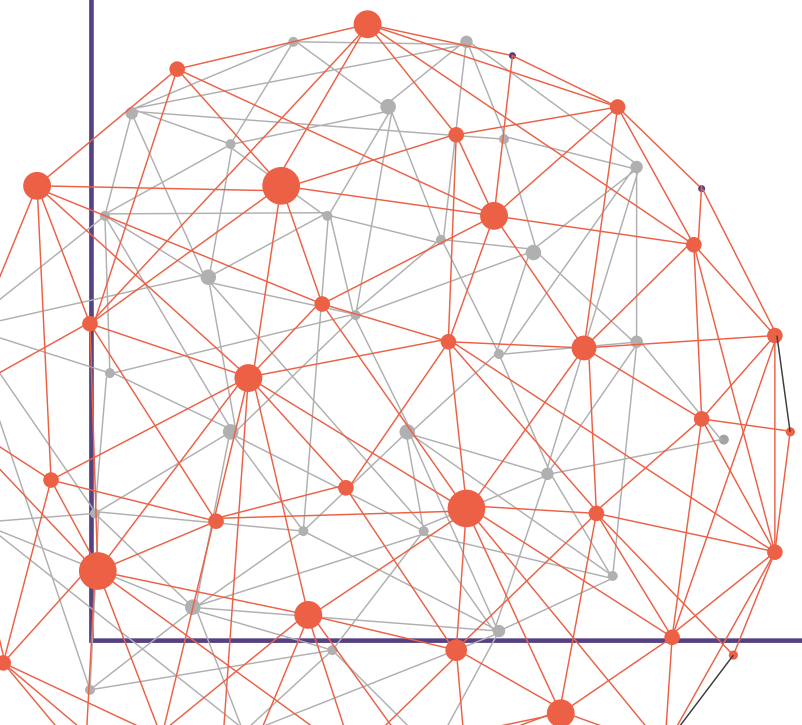
We recommend thinking of your documentation work as similar to a "digital ethnography". An ethnography is generally a form of qualitative research that involves a researcher immersing themselves in a particular community or organization to observe behavior and interactions up close. It often relies on the researcher participating in the setting or with the people being studied, at least in some marginal role, and seeking to document, in detail, patterns of social interaction in their local contexts. Digital ethnography is the online equivalent - but it is complicated somewhat by the fact that there is likely no "community" that you can identify and stay within, but a loosely connected network of people who engage with each other and share each other's content. It may help to see the online space, i.e. Facebook and YouTube, as similar to a physical space you can move around in and find people through.

Once you have gotten started (confer page 11), we recommend following accounts, clicking yourself through the comments, and looking at the accounts of frequent commentators, looking for people that the account frequently appears with or collaborates with. Let your curiosity guide you and follow leads down into the rabbit-hole. While your goal is probably just to document hate speech - and not answer wider questions about the causes of hate - letting your curiosity guide you will both help train your algorithm to show you similar content, and help you come across content in places you may not have expected it.

When trying to find new content and training your algorithm, you will have to make a choice about the degree to which you want to become involved or visible. Do you want to:

- React to posts or videos, or just screenshot them for documentation purposes?
- Try to access private groups, or only focus on public groups? (Facebook)
- Send friend requests to people with 'locked' profiles that you saw commenting on hateful posts, or ignore them? (Facebook)
- Subscribe to channels, or just save them in a list? (YouTube)

Each has implications for your own safety, ethics, as well as your ability to find content that you can verify is hate speech. On YouTube, channels generally receive less information about their viewers, while Facebook will allow the poster to see more information about you.



TERMINOLOGY

Before we proceed, we want to clarify some terminology, both to make our guide easier to read, and to clarify what kind of guidance we are offering exactly.

"Documenting" vs. "investigating" vs. "monitoring"

We are primarily providing a guide for documenting hate speech. By "documenting", we mean meticulously recording images, text, videos and supplementary information, to elucidate the statement that hate speech exists on social media. It is similar to bearing witness. We distinguish documenting from "monitoring" and "investigating". Monitoring involves explicitly setting up mechanisms to identify new hate speech whenever it appears, which often requires specific licenses or software. Finally, while you may, especially in the initial period, feel like you are investigating hate speech on social media, following leads and identifying networks of accounts that interact with each other, it is unlikely that you will be able to go into much depth and come to investigative conclusions without special tools. We therefore do not provide a guide to investigating, nor to monitoring, but to *documenting*.

"Information" vs. "evidence"

Information refers to details that are established as factual. Information need not yet have been contextualized or used to construct a narrative or an argument. Evidence is what information can turn into if it is used to support or reject a hypothesis. For instance, a piece of information that person X was in place A can turn into *evidence* to prove or disprove the specific claim that they were guilty of a crime at that place.

TERMINOLOGY
(CONTINUED)

"Data" vs. "content"

Content refers to any object that is created to be transported through the social media platform and meant to be engaged with by other users. This means that while Facebook's 'settings' section is not content, all uploaded items such as advertisements, videos, posts, events, and more are. Such content turns into data in the context of you, as a documenter or researcher, selecting particular content in order to make conclusions about them.

"Reporting" vs. "flagging"

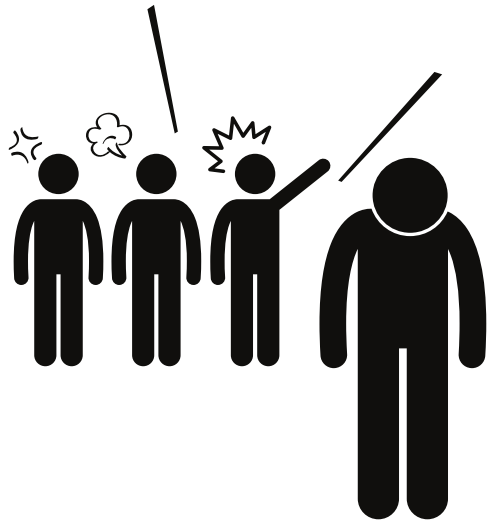
While flagging and reporting are often used interchangeably, they refer to slightly different things. Flagging is similar to alerting - when you flag content, you let the platform know that you think it is worth taking a look at, but do not necessarily specify the issue. Other platforms allow users to flag content to other users, rather than necessarily reporting it to the platform, whose intervention is reserved for rare cases. Reporting content means involving the platform itself and specifically pointing out, with reference to the platform's own rules, the reason why you disagree with the content being on the platform.

"Illegal" vs. "prohibited" content

By "prohibited" content, we refer to content that the platform itself has declared as impermissible on its own platform. This content need not necessarily be illegal under the country's law - for instance, a platform created explicitly for a political group may prohibit content supporting other political opinions. As a private platform, this is its right. However, when content is prohibited, it may sometimes also be *illegal*. In this case, it would violate a law of the country in which the user posted the content from. For instance, pornographic images or calls to genocide are prohibited under platform guidelines *and* illegal under most domestic law. For this guide, it is important to note that the aim is not necessarily to document *illegal* or *criminal* speech, but content that falls under "hate speech" under the platform's own definition.

Facebook regularly updates its definition of hate speech based on stakeholder consultations. In documenting hate speech, the key strategy is to look at the definition of hate speech applicable on the day the content was posted. The basic component defines hate speech as:

"a **direct attack** against people [...] on the basis of [...] **protected characteristics**: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define **attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation**. We also prohibit the use of harmful stereotypes, which we define as dehumanizing comparisons that have historically been used to attack, intimidate, or exclude specific groups, and that are often linked with offline violence. [...]"



Facebook prohibits content that falls within its three tier policy:

Tier 1

content targeting a person or group of people on the basis of protected characteristics or immigration status, such as:

- Violent speech or support in written or visual form
- Dehumanizing speech or imagery in the form of comparisons, generalizations, or unqualified behavioral statements (in written or visual form)
- Mocking the concept, events or victims of hate crimes even if no real person is depicted in an image.
- Designated dehumanizing comparisons, generalizations, or behavioral statements (in written or visual form)

Tier 2

attacks targeting such individuals or groups of people with statements of physical, mental or moral deficiency, or of disgust or contempt.

Tier 3

attacks which are calls to exclude or segregate a person or groups of people tantamount to expulsion or political, economic or social exclusion.

YouTube does not allow harmful, dangerous, violent and graphic content, cyberbullying and harassment, the representation of violent criminal organizations, and hate speech:

Don't post content on YouTube if the **purpose of that content** is to do one or more of the following:

- Encourage violence against individuals or groups based on any of the following attributes: Age, Caste, Disability, Ethnicity, Gender Identity and Expression, Nationality, Race, Immigration Status, Religion, Sex/Gender, Sexual Orientation, Victims of a major violent event and their kin, Veteran Status.
- Incite hatred against individuals or groups based on any of the attributes noted above.

Specific examples of content that violate this policy continues to change, but includes, for example:

- Claims that individuals or groups are physically or mentally inferior, deficient, or diseased based on any of the attributes noted above. Dehumanizing individuals or groups by calling them subhuman, comparing them to animals, insects, pests, disease, or any other non-human entity. .
- Praise or glorify violence against individuals or groups based on the attributes noted above.
- Conspiracy theories saying individuals or groups are evil, corrupt, or malicious based on any of the attributes noted above.
- Music videos promoting hateful supremacism in the lyrics, metadata, or imagery.

IDENTIFYING HATE SPEECH

Because of the lack of coherence of legal and regulatory definitions of hate speech, we typically adopt the definition of the platform we are concerned with. Even then, making a value judgment on whether content amounts to hate speech is rarely clear-cut. Speech is fluid and constantly changes, and you may not be able to accurately recognize hate speech in even another dialect of your language. Before reporting content as hate speech, go through the following steps:

Revisit the concrete definition of the platform applicable on the day content was posted. You can consider printing it and pinning it to your desk. The platform will also show a shortened version of its definition when you move to report content. Comments under content that qualifies as hate speech may also be prohibited if they express support for the content, even if they do not use notably harsh language. For instance, **a person writing "I agree" underneath a call for murder of LGBTQ persons can qualify as "support on written form".**

Get a **second opinion** from a fellow researcher and/or someone with specific **cultural knowledge**, especially if you are not familiar with the context in which the content was made. Frequently discuss the content you and your fellow researchers are marking as hate speech to ensure that you are using the same criteria and thereby establishing a reliable and valid dataset.

Consider whether the content may actually be **satire**. For this, it can help to look at:

- ***The history of the account itself.*** Does the account state anywhere that it is a satire page? Has the author been mentioned elsewhere online, such as in the news, and if yes, in what context? What other content has the account posted? In what context would they likely be using the content?
- ***Reactions to the content.*** What type of responses does the content elicit? How are comments framed, and who comments?
- ***The source of the content.*** If the content is copied or shared from elsewhere (which can be recognizable by an artist name or source on an image), what can you say about the original source?

Consider the following content:
A is a known comedian, and creates a video making fun of people of A's own religion. X shares this video, stating that they agree with it. The day before, X shared disinformation on people of this religion, stating they spread COVID intentionally. Another day before, X posted an opinion piece about the threat people of that religion allegedly pose to society.

FACEBOOK PRACTICE CONTENT: IS THIS HATE SPEECH?

This is a practice exercise. The context to these is missing, and you therefore do not have all the information needed to make this value judgment. Think about what further information you might need to identify whether this is hate speech. The content may also violate platform standards on other grounds, but consider here whether it amounts to hate speech.

☑ Krk Reddy
Massacre this pig antinationalist traitor and badmaash badkav bastard in public
Like · 1w

☐ Aparna Das
They were created through violence, nothing good can't be expected from those jihadis
Like · 12h

☐ 

☐ 

☑ Beenanambiar Beena
Barks is a barking mad dog who licks the ass of Pak for .oney she is a dirty gutter snipe a bloody basteted
Like · 1w

☐ Krishnadev Goswami
Barkha is a hallmark Bitch!!
Like · 1w

☑ Sharad Varik
B D is a rotten bloody C**t , without any doubt ! Never ever go near her !!
Like · 1w

☐ 

☐ 

HOW TO DEAL WITH BORDERLINE CONTENT?

In 2018, Mark Zuckerburg, while discussing bordering content, wrote:

"When left unchecked, people will engage disproportionately with more sensationalist and provocative content. Our research suggests that no matter where we draw the lines for what is allowed, as a piece of content gets close to that line, people will engage with it more on average – even when they tell us afterwards, they don't like the content."

Facebook then adopted recalibration of borderline content to reduce virality. However, it is unclear how and when Facebook's algorithms classify content as "borderline content".

To help resolve the issue in reporting borderline content, we advise using the following parameters:

- Does the content use terminology in the form of ridicule or slang? For example, the word 'sikular' is used to denounce secular values. Or calling Muslims 'peaceful' when the accompanying post itself in essence ridicules them?
- Who is the author of the content? Does the author regularly post content classified as hate speech, ridicule, incitement to violence? Or does the author regularly support hateful and incitement to violence narrative?
- Can you determine the context within which the author made the online remark?
- Is there a particular time period when similar content/narrative gained virality across social media? But such virality was creating a negative image of minorities in the region you are concerned with?

With borderline content, we tend to follow the same steps we would for clear violations: That is, we flag the content as hate speech or incitement to violence. Given that Facebook does not have any means through which users can flag borderline content, we also make multiple reports under multiple closely related categories, including false news.

WHAT DO YOU NEED?

We want the methodology we present to be accessible to everyone who is motivated to document hate speech on social media. In order to follow the steps in this guidebook, you therefore only need:

- Access to internet on any device
- A social media account (on Facebook, YouTube/Google). We recommend making a second account for documentation purposes (cf. Section "Security").

It can help to have access to Open Source Intelligence (OSINT), using such techniques requires additional security steps and knowledge of more complex methods. It can also help to use CrowdTangle, an analytic tool to monitor keywords, tags and more on social media. CrowdTangle access is available for specific Facebook partners, such as media publishers and public figures, as well as academics and researchers in specific fields. **While using OSINT and tools such as CrowdTangle could therefore seriously assist you, you may very likely not have access to them**.

SECURITY

Before starting to document online hate speech, take steps to improve your privacy and security. This is especially important if you yourself belong to a group with protected characteristics, as you are therefore especially vulnerable to being targeted by the authors of the content you are documenting. Even if you do not think that there is a risk of negative consequences for you, taking security steps is essential in general to safeguard other data you may have, as well as your fellow researchers.

Create a dedicated account for documentation purposes.

You can use Gmail, or this temporary email generator: https://10minutemail.com/

Do not use Chrome as your browser. Preferably delete the browser from your device altogether, and switch to:

- *DuckDuckGo* instead of the Facebook app (on your phone)
- *Brave* (on your computer)

Install a VPN (virtual private network) on the devices you will be using. A VPN creates an access point to the internet that hides your digital identity, and is completely legal. If you log onto an account with your details, though, using a VPN will be VPN useless. We recommend *NordVPN*.

Only use a private or password-protected wifi for sensitive work. Public open wifi can easily be hacked.

Using a middleware can be helpful. We often use archive.is extension to archive posts and pages which were particularly toxic in nature. Other options can include Evernote and Archive Box.

For saving documentation or sharing notes with fellow researchers use cryptdrive, or OneDrive as more secure options

SECURITY: SOFT SKILLS

It makes sense to work in a team for many reasons, including ease, efficacy, and having a second opinion. However, be aware of the extent to which you can trust strangers on the internet. While scaling up the research and documentation process can feel empowering, it can invite unwanted attention from people who do not share your aim. Before sharing information with a research partner, ask yourself whether you can trust them and if their understanding of the issue is same as yours. We advise building rapport before working together.

While you are browsing social media, you may stumble upon groups or profiles that are hidden to the public (i.e. 'locked' profiles or private groups on Facebook). While you are free to send connection requests to those profiles, and can try to gain access to the private groups, be aware that this may have implications for your own security.

GETTING STARTED

In general, algorithms on social media help move you towards similar content to the content you've already positively reacted to. This means that once you have engaged with hateful content, you're likely to discover more - but how do you get out of your own content bubble in the first place, and into the new one?

When you make the step from reporting just one post to starting a documentation process, we strongly recommend making a second account. Make sure not to violate the platform's community standards when setting up that account (i.e. impersonating someone).

Scenario 1:
You do not yet have a lead.

Option 1: Keyword Search

You can start finding content by doing a keyword search on the platform. The keywords will vary based on the type of hate speech. Before you start, ask yourself what kinds of hate speech you want to identify. What hate speech do you feel confident in distinguishing from satire? Who have you recently seen being targeted by hate speech? When deciding on keywords, make sure to do your background research and to not let stereotypes guide you. Racial slurs or hateful language you have seen on TV will likely not be in use anymore, or only lead you to satirical posts. Keywords mentioned in the platforms' community standards are also likely not up to date.

Option 2: Identifying actors

Facebook, YouTube and other platforms occasionally remove actors from their platforms *en masse*. You can find press releases for this and try to identify real people behind the profiles/pages/channels. This may be directly mentioned in the press release, or require you looking for a website with the same name as the deleted profile/page/channel.

However, when finding actors you think may be posting hate speech, be wary of your pre-conceived biases about the people you think are posting hate speech: Just because a person holds a certain opinion does not mean that they will also be propagating it through hate speech.

Option 3: Fact-checks

Debunked news and fact-checks are good starting points for identifying content. Some pages that provide such fact-checks are: Alt News (India), Factuel (France), CORRECTIV (Germany), Newtral (Spain), BBC Reality Check (UK), AFP Fact Check (North America), more here. This can help you find actors you can look up on social media, identify keywords that are currently being used to dehumanize groups with protected characteristics, etc.

Then move to scenario 3.

Note: The purpose of finding content is to document it, and through it find more content. The focus of your documentation is initially the content itself, rather than the actors posting it (see Section "Using the Data").

Scenario 2:
You already found hateful content elsewhere.

Scenario 2.1: Post on other social media

If you have found content on Twitter or elsewhere, for instance for reason of it being viral, you can try to find it on other platforms by searching for keywords related to it. If you find a hateful video on YouTube, you can see if people have engaged with it on Facebook by copying the name of the video and pasting it in the search bar on Facebook.

Scenario 2.2.: Website

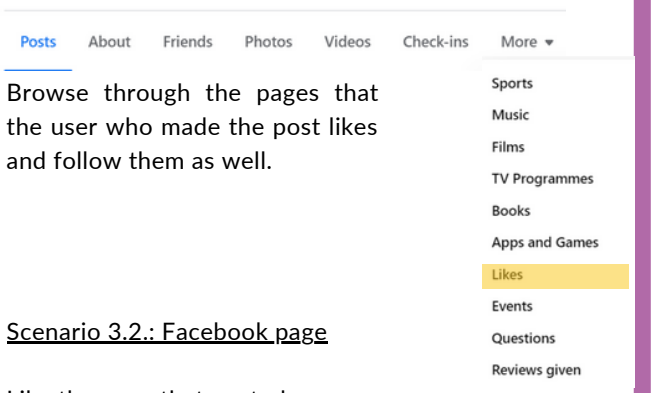
If you have found content on a website, for instance a hateful article, you can try to find how it was shared by copying the name or URL of the article and pasting it in the search bar on Facebook. If an author name is given, you can try to find them on the platform. If no authors or names are given, you can put the website's URL in *https://who.is* and with luck learn who the website is registered to.

Then move to scenario 3.

Scenario 3:
You already found hateful content on the platform.

If you have already seen hateful content on Facebook or YouTube, you can continue from there and train your algorithm to show you more similar content.

Scenario 3.1.: Facebook private profile



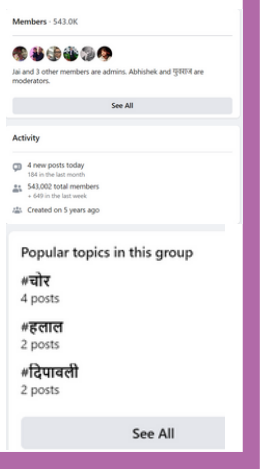
Browse through the pages that the user who made the post likes and follow them as well.

Scenario 3.2.: Facebook page

Like the page that posted the content.

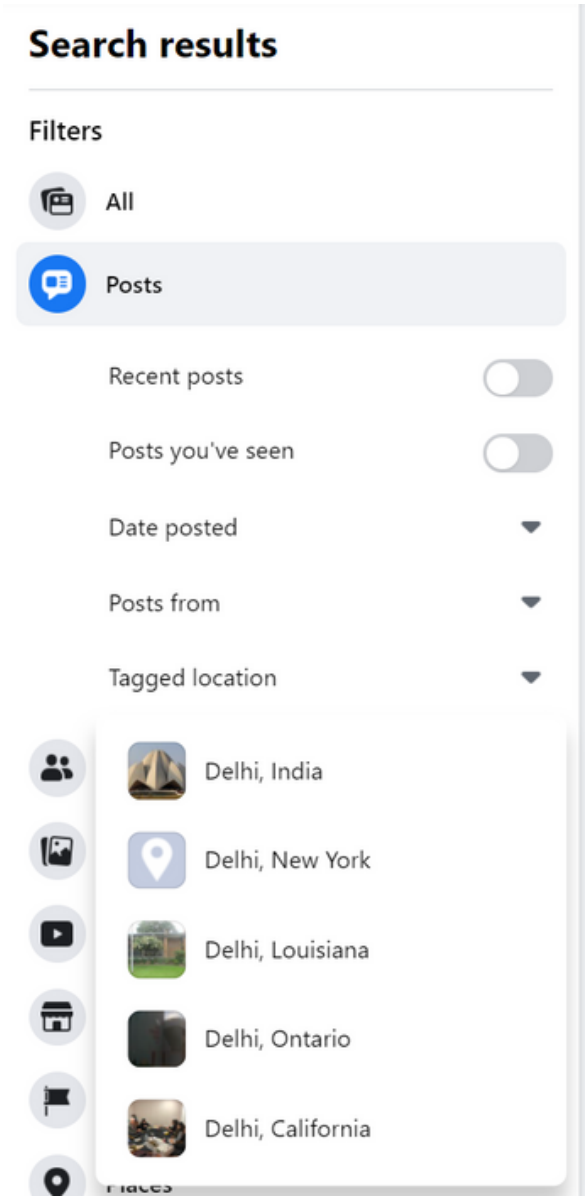
Scenario 3.3. Facebook group

If the post was in a public group, see if the group has popular topics related to the hateful post. If yes, you can consider joining the group if you feel comfortable. Identify whether the group is very active, has enough members so it would merit your attention, and take a look who its moderators are.

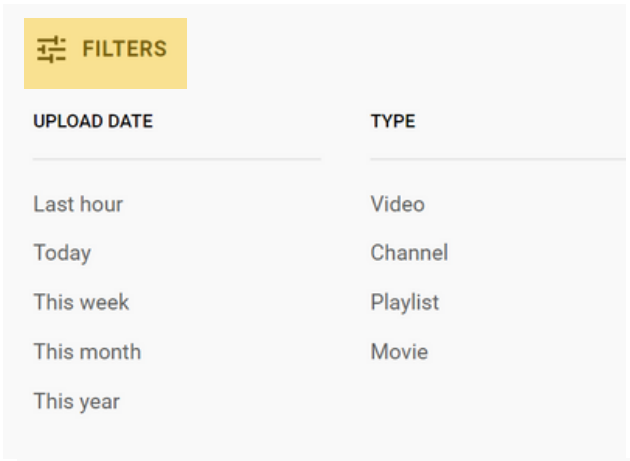


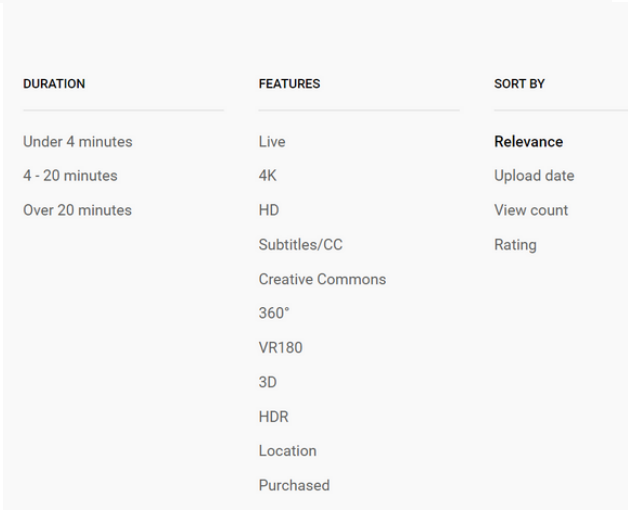
FINDING EXACTLY WHAT YOU'RE LOOKING FOR

Refining your search criteria and telling the search bar exactly what you are looking for can be tricky. While YouTube allows so-called "Boolean operators" as well as filters, search results on Facebook have to be filtered via Facebook's own levers. Here are some instructions for getting more precise results:



Looking for posts by location can help identify posts that were concern a certain event that featured in the news, e.g. a political rally at which hate speech was raised.





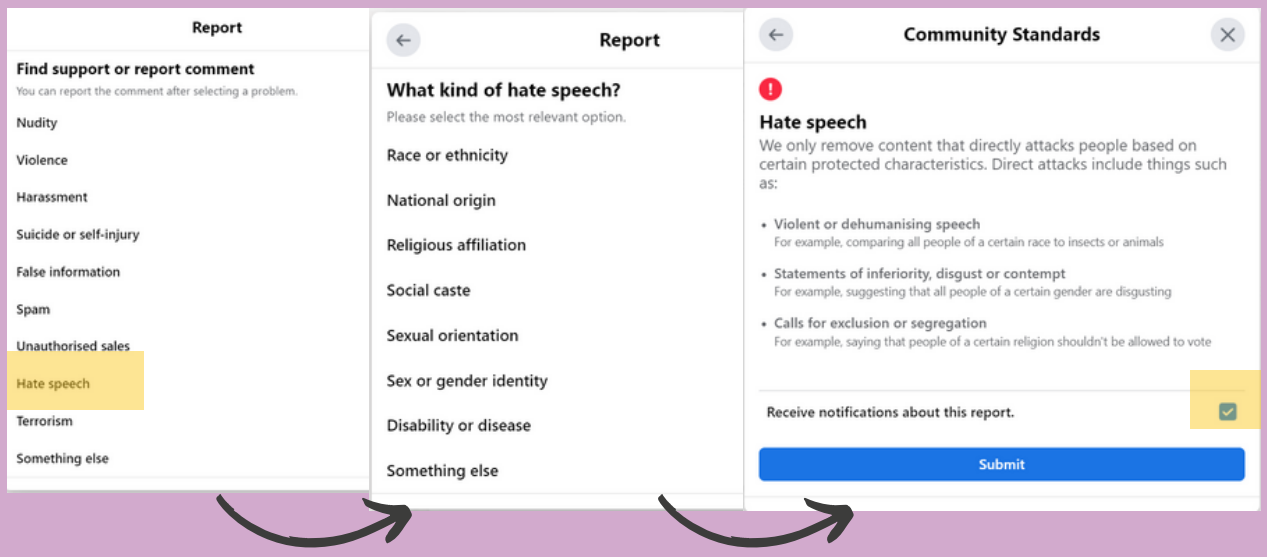
Search by	Search Syntax
Search only in title	intitle:search query
Force search for word in results	+word
Omit a word from results	-word
Search for exact composition of words	"search query"
Wild character for word	*

REPORTING CONTENT

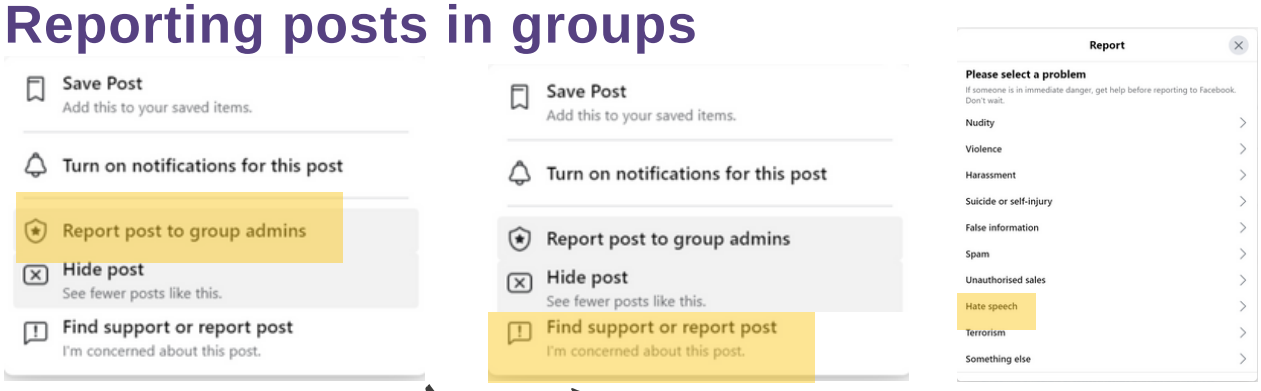
An essential part of your documentation work is reporting content to the platform for violating its own standards and guidelines, and documenting the response received from the platform. Reporting procedures require a very rigid classification of content, as it is not possible to report intersectionally harmful content under several headings at the same time. We will show here how to report *hate speech* - of course, you may encounter content that is rather "violence" or "harassment", and you should then follow those options.

facebook

Reporting comments



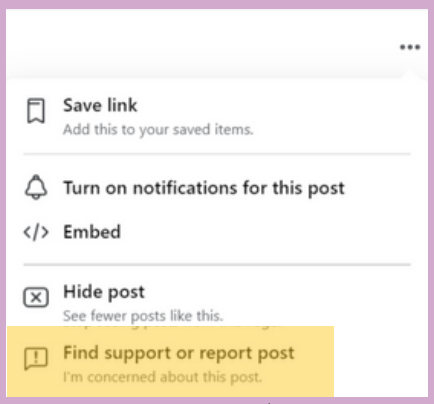
Reporting posts in groups

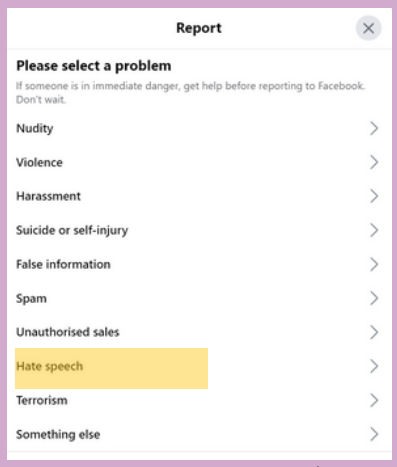


Before reporting a post in a group to Facebook, report the post to the admins, as they have the power to moderate the group's content. and see what response you receive from them. If they do not contest the content, report it to Facebook.

Same procedure as above.

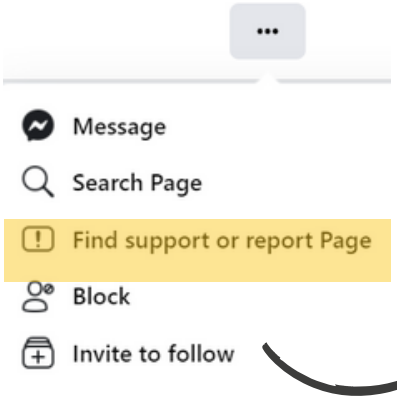
Reporting posts

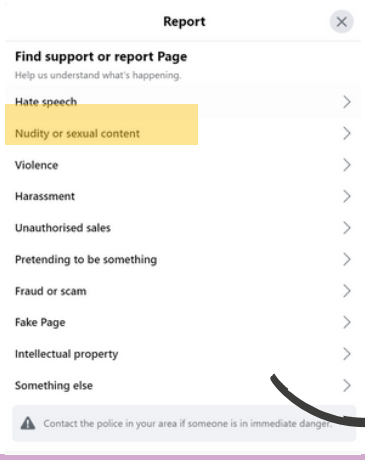




Same procedure as above.

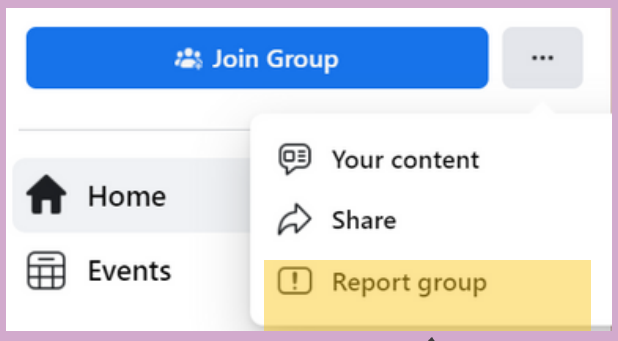
Reporting pages

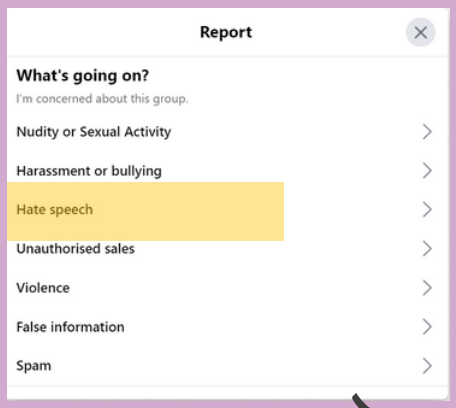




Same procedure as above.

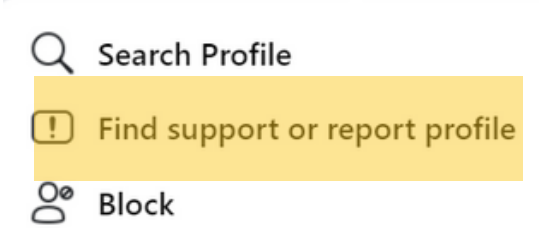
Reporting groups

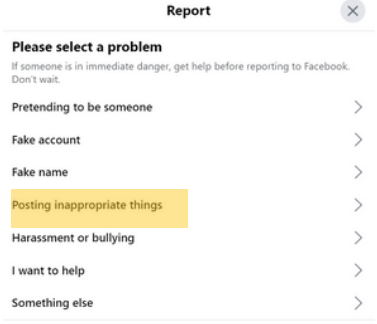




Same procedure as above.

Reporting profiles

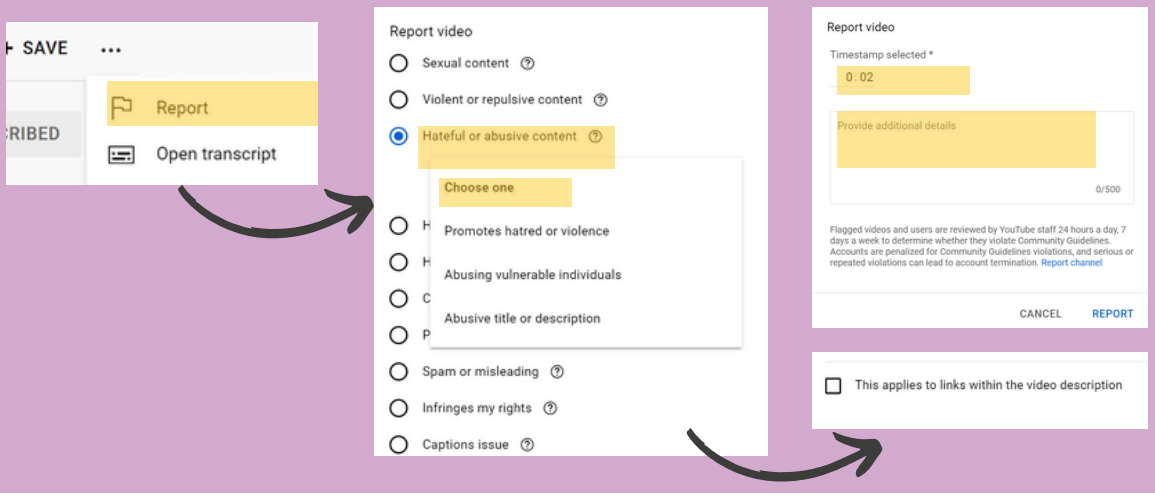




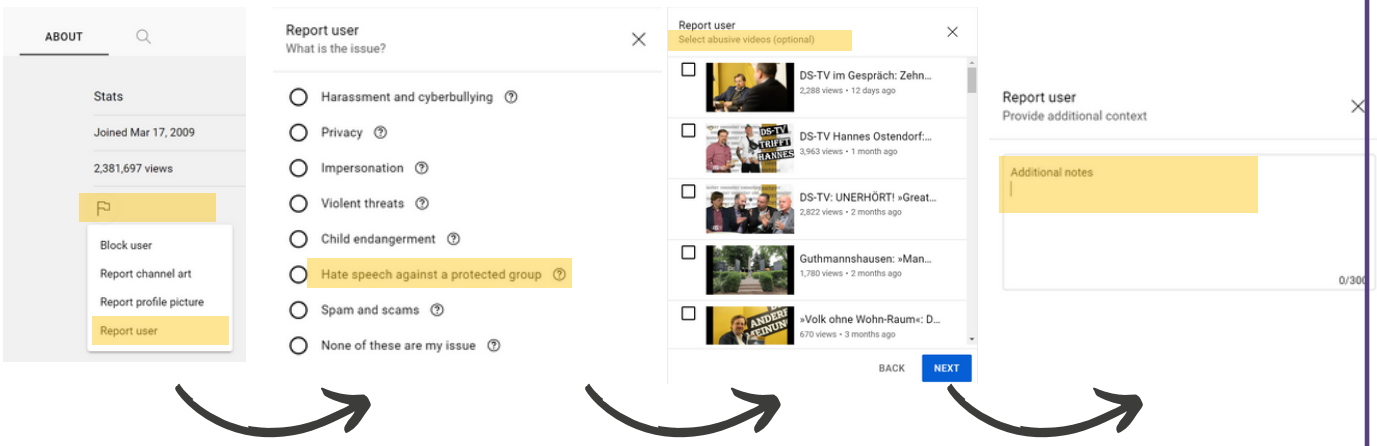
Facebook is more concerned with fake profiles or names. Before reporting any individual, be certain that they are completely abusing Facebook for propagating hate.

YouTube allows users to not only report content, but to give explanations and to highlight linked content in the video description. If it is not the title itself that is hateful or abusive, YouTube requires users to select a specific timestamp in the video. YouTube details its process here.

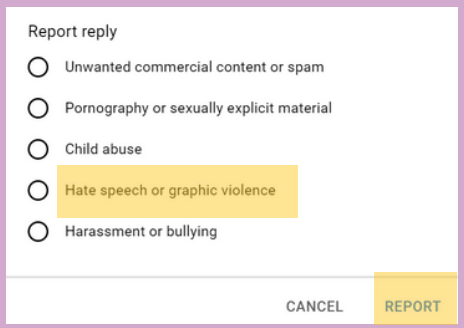
Reporting videos as hate speech



Reporting channels as hate speech



Reporting comments as hate speech



THOROUGH DOCUMENTATION

The content you identify can become relevant long after. Thoroughly documenting not just the content you saw, but where you saw it, and how people interacted with it is therefore essential. It is important to record as much information as you can immediately, as the content might be deleted later on and become inaccessible to you.

To archive the content properly, you should try to get the following:

1

A hard copy

Start by taking a screenshot of the content.

For comments, this will be the most you can do. Try getting a screenshot of relevant comments in context of the content they are commenting on.

For videos, download the video via https://en.savefrom.net/55/ or a similar online video downloading site. Make sure to block your ads, and to not use illegal file sharing sites.

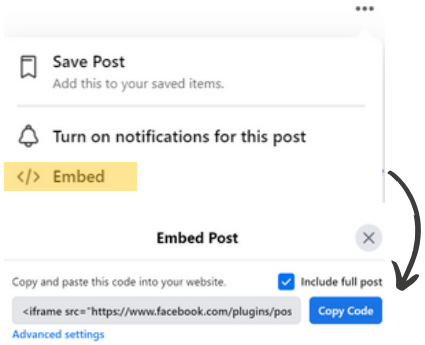
2

An archived link that will link you to the current content even if it is removed or changed

Copy the link of the content, go to https://archive.org/ and "save page now". The link you will receive is an archived link.

Finding the link to Facebook posts can sometimes be difficult if you come across it in your feed. Copy the "embed" link and delete everything starting at *&show_text=true*

Comments under Facebook posts and YouTube videos do not have links, so you must rely on only screenshots here.



DUBLIN CORE METADATA

The screenshot of the content is your data and hard documentation. It is similar to taking a picture at a crime scene. In order to make sense of the data, though, you need metadata - data about your data. We recommend documenting as much as you can from the *Dublin Core Metadata Element Set*, a set of fifteen core properties that are in standardized use globally. These are:

1. **Subject** – "The topic of the resource".
2. **Title** – "A name given to the resource".
3. **Type** – "The nature or genre of the resource".
4. **Format** – "The file format, physical medium, or dimensions of the resource".
5. **Creator** – "An entity primarily responsible for making the resource".
6. **Contributor** – "An entity responsible for making contributions to the resource".
7. **Publisher** – "An entity responsible for making the resource available".
8. **Date** – "A point or period of time associated with an event in the lifecycle of the resource".
9. **Source** – "A related resource from which the described resource is derived".
10. **Coverage** – "The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant".
11. **Description** – "An account of the resource".
12. **Identifier** – "An unambiguous reference to the resource within a given context".
13. **Language** – "A language of the resource".
14. **Relation** – "A related resource".
15. **Rights** – "Information about rights held in and over the resource".

Available/applicable metadata for each piece of data will vary. You can see most of it directly, and can right-click on social media content and select "inspect" in order to see more. A suggestion for easy documentation is available on the following page:

Documentation form template

Screenshot of content

Archived URL

Name of item (if applicable)

Name of page/channel/group/profile

First creator

Number of shares (if applicable)

Number of views (if applicable)

Number of comments (if applicable)

Number of likes (if applicable)

Location tagged (if applicable)

Keywords on type of hate speech

Type (video, image, ...)

Platform

Language

Date posted

Description

Contextual information

Reported to the platform as/ on

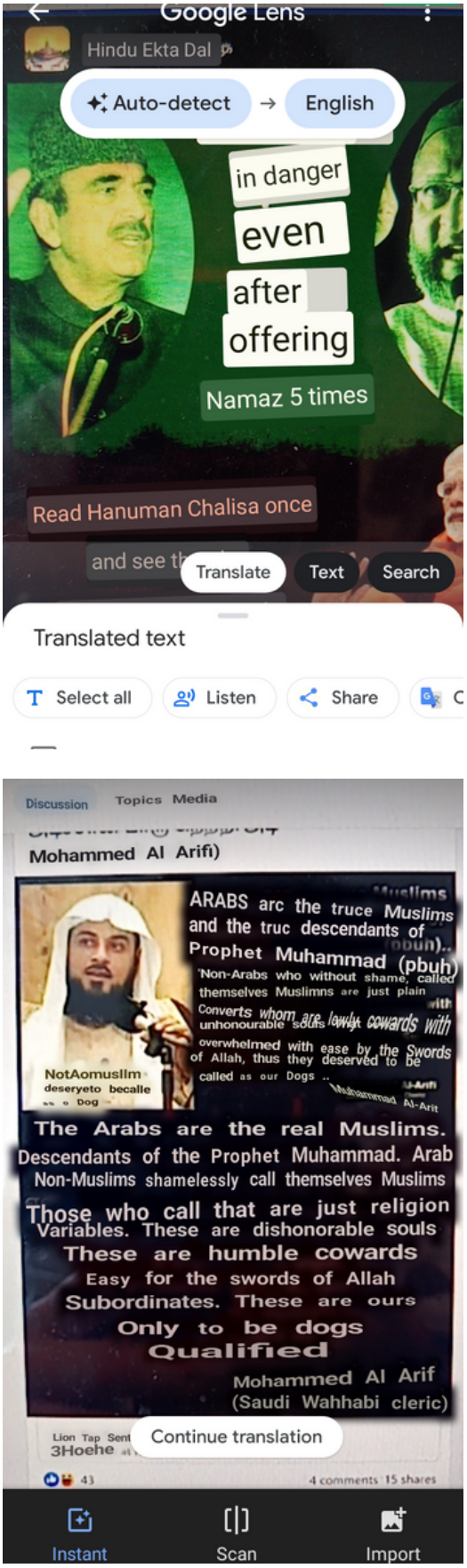
Response from the platform

https://forms.office.com/r/wQgUHEQh6D

HANDLING DIFFERENT LANGUAGES

Some social media platforms automatically translate content into the language you are using the platform in. This is the case not only on Facebook, where text posts are shown in your language as well as the original language, but also on YouTube, in which video titles are sometimes automatically translated into your language without you realizing. This is rather the case with large YouTube channels, though, and will likely not be the case for the content you work with. However, even in the presence of convenient automated translation, we recommend to only consider content in a language that you are confident in. Facebook, for instance, claims that it trains its algorithm with words "based on local nuance", and even then fails to identify a majority of hate speech on its own platform. Knowing the cultural context of speech as well as being able to understand the nuances of language is therefore essential in this process.

It may nonetheless be useful to get a basic translation of content in order to add context to other content or discern whether an account is satirical or serious. For text content, you can simply copy/paste it into Google Translate or DeepL.com. If you have an image with text that you cannot copy, you can use the phone apps Google Translate or Google Lens to scan images and get immediate approximate translations. You can also take a screenshot and import it into the Google Translate app.



DEALING WITH FAKE(?) PROFILES

On Facebook, as well as in YouTube comments, you may come across users that appear dodgy. While their contributions on social media may have a real impact, they may not necessarily represent real people. This does not necessarily have to mean that they are bots, but could mean that it was created explicitly for the purpose of sharing harmful content, such as disinformation and hate speech. Indications of this are:

- The user has disproportionately many friends, or barely any at all;
- Their profile pictures are of poor quality, and/or can be easily found online with a reverse Google Images search;
- The barely have any content, and the content that they do have is hate speech, trolling or similar harmful content;
- The name in their account URL does not match the name in their profile.

Identifying whether a profile may be "fake" has implications both for whether you just report their content or their entire profile, and for how you further engage with the data you collect on the account.

We recommend reporting such profiles as coordinated inauthentic behaviour in addition to reporting the content they post under the category of hate speech, violence or false news depending on the context. If you are satisfied that the content posted on the account is harmful, we recommend reporting the profile as "posting inappropriate things" (Facebook) or the channel as "hate speech against a protected group" (YouTube). Similarly, when documenting, your first concern should be whether the account is propagating hate and an additional remark on the fakeness of the profile can also be made.

USING THE DATA

Once you have transformed the platform content into usable data, including meta data, you may wonder what you can do with it. If you have reported it to Facebook/YouTube, and they have chosen not to remove it, what further steps are there? The answer to this question absolutely depends on the motivation with which you started into this project. Here are some guiding questions to help you decide how further to use the data:

- Are you part of a larger team with research funding or institutional backing, or are you working alone? What capacities do you have?
- Are you more concerned with the platform's failure to moderate content, or with the authors of the content itself?
- Can you reach out to the platform with your dataset, and is there any chance that they will respond and engage with you?
- Who were the actors who were propagating hate speech? Did you identify any particular political party or group? Are they currently under investigation for related topics, or should authorities be informed?
- Does the content simply violate platform standards, or is it also explicitly prohibited under the law of the country it was posted from? Could your dataset hold in a court as evidence, or is it rather data of relevance for research purposes?
- Do you think an attack on a protected group is imminent?
- Do you think involving the authorities, such as law enforcement, is at all helpful, or are there issues of institutionalized racism or other discrimination that would lead to adverse consequences?

If you use our embedded form above, we will be able to refer to and analyze the content you identified in our advocacy work.

LOOKING AFTER YOURSELF

Documenting hate speech can be emotionally exhausting. You may come across videos with "content warnings" that show graphic violence, but haven't been taken down by the platform because they identified it as 'raising awareness about little discussed issues'. While you may not be 'out in the field', taking explicit, pro-active steps to take care of your own wellbeing is essential.

DAILY SELF-CARE STEPS

- Remind yourself why you are doing this work, and who you are doing it for.
- Exercise, eat well, sleep enough.
- Actively seek out good news. You can use websites like the Good News Network and Good News EU.
- At the end of the day, write down things you can be grateful for or happy about that happened during the day, or that happened in the world in general.
- Consult academic literature on hate if you want. It can help you make sense of the content you are seeing.
- Re-watch movies or series you know are comforting to you. It gives your brain rest.
- Log out of your account when you call it a day.

RECOGNIZING WARNING SIGNS

Every person responds differently and to different triggers. Here are some common warning signs:

- Feeling agitated at minor things, such as friends and colleagues asking for clarifications
- Feeling immobile, exhausted, or low on energy
- Crying while documenting content*
- Clenched fists, biting nails, pacing, or similar body-focused stress signs
- Feeling no capacity to engage with problems of others

**While crying is a normal emotional response to hate speech and similar content, it is a sign that you should be taking breaks and need a positive balance.*

A HEALTHY FRAMEWORK

Further above, we have emphasized the importance of working in a team for practical reasons. We cannot understate the value of working in a team for your own mental wellbeing. As a team, find a way to do this work that can help you cope with exhausting material. Inform your team members of what content may particularly serve as a trigger. Set up a schedule with 'working hours' to avoid encountering hate speech or worse before going to bed. Explicitly dedicate parts of your conversations to checking in on each other, and practice active listening techniques and other tools presented here. If you are applying for budget for hate speech documentation, incorporate a wellbeing fund into your proposal.

EMERGENCY RESOURCES

Depending on the content you come across, simply taking daily self-care steps may not alleviate the emotional burden. Beyond Empathy Cafes and Empathy Circles, there are only a few resources specifically for activists:

- You can refer to the Trained Emotional Support Network (TESN) affiliated with the climate group Extinction Rebellion, and their Don't Panic team for resources on what to do in an emergency.
- The Trevor Project hotline provides confidential support for LGBTQ people under 25 in English.
- The Metabot is a chatbot available in German that gives guidance on handling discrimination.
- The Animal Activist Support Line provides support to burnt out animal rights activists.
- Click here for an interactive tool to help you find more general Mental Health Support.

A CIVIL SOCIETY GUIDE TO DOCUMENTING ONLINE HATE SPEECH

Stichting The London Story
www.thelondonstory.org