2

How to Identify and Report Hate Speech

02 HOW TO IDENTIFY & REPORT HATE SPEECH
CONTENTS PAGE

#DEFYHATENOW | CHAPTER 2

HOW TO IDENTIFY & REPORT HATE SPEECH

What is Hate Speech?

"In broad terms, hate speech is a communication that denigrates people on the basis of their membership of a particular group. This can include any form of expression, such as images, plays, and songs, as well as speech.

Some definitions even extend the concept of hate speech to include communications that foster a climate of prejudice and intolerance – the thinking here is that these kinds of communications may fuel discrimination, hostility and violent attacks later on. [DW Hate Speech FAQ]

DEFINITION OF DANGEROUS SPEECH

by Susan Benesch, Dangerous Speech Project:

"Inflammatory public speech rises steadily before outbreaks of mass violence, suggesting that it is a precursor or even a prerequisite for violence, which makes sense: groups of killers do not form spontaneously. In most cases, a few influential speakers gradually incite a group to violence.

Violence may be prevented, then, by interfering with this process in any of several ways: inhibiting the speech, limiting its dissemination, undermining the credibility of the speaker, or 'inoculating' the audience against the speech so that it is less influential, or dangerous."

DISCUSSION: WHAT ARE THE EFFECTS OF HATE SPEECH AND DANGEROUS SPEECH?

Ask everyone in the group to briefly introduce themselves and to give their understanding or definition of hate speech. Appoint someone to write down key points of these definitions on a flipchart, post-it notes or whiteboard as they are being given, to give the group a visual.

- Discuss with your group, school or workshop participants the impact of hate speech
- Introduce your experiences and understanding of hate speech and dangerous speech
- Jot down notes and terms on flip chart or post-its or chalkboard as they arise

NOTES FOR FACILITATORS

- 1. Moderators/presenters are not to engage in political opinions in regard to the current conflict
- 2. Participants should not engage in political debate or hail insults - please strive to remain neutral
- 3. Time keeper and moderators have the right to stop anyone who diverges from main topic

SOME DRIVERS FOR ONLINE HATE SPEECH TO CONSIDER IN YOUR DISCUSSION

- Children are affected by or involved in violence at increasingly earlier ages - due to proximity to conflict, learning to hate, breakdown of positive norms & values.
- Youth feel frustration, social media offers an open platform, which can further entrench attitudes of hate.
- Lack of accountability - no reconciliation process, the power of anonymity and geographic distance all contribute to a lack of consequence for online incitement.
- Political & tribal alignments - misconceptions on cultural diversity and conflict narratives based on tribal affiliations
- Lack of policies to balance freedom of expression, ethics and privacy against curbing hate speech online
- Online reactions amplify tensions between citizens/groups increasingly splintering communities further, even within diaspora.
- Vicious circle of increasing brutality, dangerous and aggressive speech and directed incitement moving between inaccurate reporting of the conflict to de-contextualisation in social media.
- Increasing cases of networking between members of the South Sudanese diaspora (e.g. in USA and Australia) using social media platforms to organise campaigns of directed incitement on the ground in South Sudan.

Links between social media usage and offline activity are reflected in-country: e.g. messages travel from Facebook to mobile phones to radio, graffiti and word of mouth.

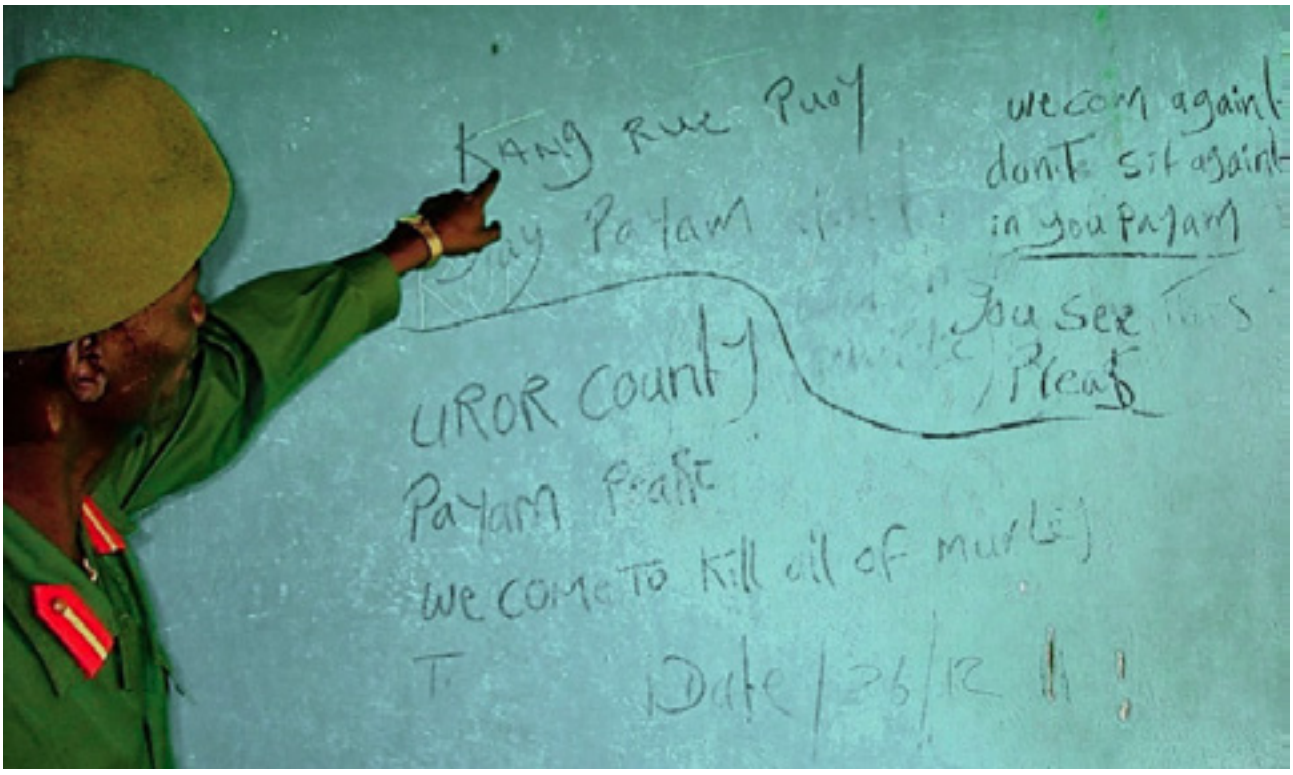

Photo by Ferdinand von Habsburg-Lothringen

PLAY VIDEO [USB STICK]

SOCIAL MEDIA AS A WEAPON OF WAR

HATE SPEECH IN SOUTH SUDAN
https://youtu.be/6TpEF721Gh8
(Short version 03'43")

PeaceTech Lab works to reduce violent conflict using technology, media, and data to accelerate and scale peacebuilding efforts.
In 2016, PeaceTech Lab conducted research to better understand the connection between online hate speech and violence on the ground in South Sudan. Learn more about the project here: **http://www.peacetechlab.org/hate-speech-in-south-sudan/**

HATE SPEECH GUIDELINES & DEFINITIONS

The International Covenant on Civil and Political Rights (ICCPR), a UN treaty, calls on governments to prevent hate speech. Article 20(2) of the ICCPR says: "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law."

Hate speech laws are a relatively modern phenomenon that appeared in Europe in the wake of World War II. The idea behind such laws was to curb the kinds of anti-Semitic and racist propaganda that gave rise to the Holocaust. Germany, Poland, Hungary and Austria passed hate legislation decades ago.

Many other countries have since followed suit. For example, under Kenyan law, a person commits an offense if they stir up "ethnic hatred". France goes further. Its laws forbid any communication intended to incite discrimination, hatred or harm regarding ethnicity, nation, race, religion, sex, sexual orientation, or handicap.

What about online hate speech?
Hate speech often shows up online, especially on social media. Facebook, Twitter and Google each has its own specific definition of hate speech and their approaches to dealing with it are evolving.

The **UN's International Committee on the Elimination of Racial Discrimination** understands 'hate speech' as "a form of other-directed speech which rejects the core human rights principles of human dignity and equality and seeks to degrade the standing of individuals and groups in the estimation of society."

The Broadcasting Complaints Commission of South Africa considers 'hate speech' to be "material which, judged within context sanctions, promotes or glamorizes violence based on race, national or ethnic origin, colour, religion, gender, sexual orientation, age, or mental or physical disability" or "propaganda for war; incitement of imminent violence; or advocacy of hatred that is based on race, ethnicity, gender or religion, and that constitutes incitement to cause harm."

The **European Court of Human Rights**, in a definition adopted by the Council of Europe's Committee of Ministers, considers 'hate speech' as: "all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility towards minorities, migrants and people of immigrant origin."

What about freedom of expression?

A tricky question. Critics of hate speech laws say such laws infringe upon freedom of expression. Others say other rights, such as equality and freedom from discrimination and violence are also important. It's an ongoing debate that shows how hard it is to balance the right to voicing opinions with protecting community interests and deterring hate crimes.

Some argue that if you suppress free speech, you just push it underground and it doesn't go away. In fact, it can become more dangerous there, because it's not in the public sphere. Others say that one protected freedom, namely speech, cannot always trump other rights, such as equality and freedom from discrimination. They add that repeated expression of hateful messages can gain social traction.

What are some of the effects of hate speech?

There are real-world examples of hate speech having disastrous, deadly results. Nazi Germany is a prime example. The viciously anti-Semitic newspaper Der Stürmer energetically encouraged the German people to persecute and even exterminate Jews. The Nuremberg Tribunal ruled against the paper's publisher, Julius Streicher, holding that incitement to genocide is a crime under international law.

More recently, the 1994 genocide in Rwanda is an example where it is widely believed that hate speech played a significant role in the massacre of 800,000 Tutsis and Hutus. A private radio station there called for people to "exterminate the cockroaches", even broadcasting lists of people to be killed and telling killers where to find them.

In the aftermath of the December 2007 presidential elections in Kenya, violence erupted, mainly between Kenya's three largest ethnic groups. More than 1,100 people were killed. A popular radio broadcaster, Joshua Arap Sang, was accused of using his position to encourage ethnic attacks. Text messages were widely circulated calling on one group or another to "exterminate" ethnic rivals. Since then, Kenya has passed laws prohibiting hate speech.

PEACETECH LAB SOCIAL MEDIA & CONFLICT IN SOUTH SUDAN LEXICON OF HATE SPEECH TERMS
PeaceTech Lab developed the Social Media Lexicon of Hate Speech Terms, combining cutting-edge social media analysis with in-country expertise to identify both the terms likely to incite violence, and their social and political context.

The Lexicon identifies alternative language that would mitigate the impact of this speech. The goal is to inform organisations and individuals combating hate speech and building peace in South Sudan, and raise awareness among social media users on the dangers of specific inflammatory language.

The PeaceTech Lab lexicon can help facilitate a discussion about the specific context of South Sudan, using examples of hate speech terms from these reports. [SEE Handouts & Exercises 2.]
http://www.peacetechlab.org/hate-speech-in-south-sudan/

CATEGORIES & EXAMPLES OF HATE SPEECH ONLINE

Discuss the context and intention of examples of hate speech
[See A2 poster & exercise in handouts]

Look at the images that show different types of hate speech or dangerous speech.
You can also select and show your own examples if you have experienced this personally.
Show examples of when you have seen people sharing online hate or dangerous speech inciting violence.

Use this material and the poster to discuss the various types of hate speech and discuss how is hate speech different from dangerous speech? Outline what makes dangerous speech and incitement – there is a clear call to action – it does not always include hate speech.

3 CATEGORIES OF ONLINE INCITEMENT

1. Emotional: Reinforcing negative stereotypes & 'other.' Lack of social media ethics; lack of understanding of the consequences and effects of online activity.
2. Personal or group virulence: De-humanising the other, propaganda, image/fact manipulation.
3. Organised, aimed, directed Incitement to violence: Hoax / deliberate spread of rumours to spark violence or armed action, potential for genocide.

DISCUSS YOUR OWN EXPERIENCES OF HATE SPEECH OR DANGEROUS SPEECH ONLINE AND IN YOUR COMMUNITY

- Who are the targets of hate speech in the examples on the poster?
- How might hate speech affect the people who are targeted?
- What consequences might these examples of hate speech have on people identifying with the communities targeted?
- What effect does it have on society in general?
- How does listening to hate speech or dangerous speech on the radio, in person or online make you feel?
- What could you do to stop the spread of hate speech in your own community, your family or school?

HATE SPEECH GUIDELINES ON FACEBOOK

https://www.facebook.com/communitystandards#hate-speech

"Facebook removes hate speech, which includes content that directly attacks people based on their: Race, Ethnicity, National origin, Religious affiliation, Sexual orientation, Sex, gender, or gender identity, or Serious disabilities or diseases. We allow humor, satire, or social commentary related to these topics. Sometimes people share content containing someone else's hate speech for the purpose of raising awareness or educating others about that hate speech. We expect people to clearly indicate their purpose, which helps us better understand why they shared that content. We carefully review reports of threatening language to identify serious threats of harm to public and personal safety. We remove credible threats of physical harm to individuals."

HATE SPEECH & VIOLENCE GUIDELINES ON TWITTER

https://support.twitter.com/articles/18311

"**Hateful conduct**: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease. **Violent threats (direct or indirect)**: You may not make threats of violence or promote violence, including threatening or promoting terrorism."

VIDEO [USB STICK]

HOW TO REPORT HATE SPEECH - VIDEOS BY PEACETECHLAB

Countering Hate Speech – How to report on Facebook
https://youtu.be/_kRNx1WHAm0

Countering Hate Speech – How to report on WhatsApp
https://youtu.be/1PlHuMl0ndQ

Countering Hate Speech – How to report on Twitter
https://youtu.be/tLwgoTgbf8E

Countering Hate Speech – How to report hate speech on YouTube
https://youtu.be/sacqQNyxVe8

HOW TO REPORT ABUSE ON SOCIAL MEDIA PLATFORMS

Social media platforms been criticised for their handling of complaints. They are not always responsive to user concerns, but overall they do continue to assess new ways to monitor & counter hate speech.

HOW TO REPORT ABUSE & THREATS ON FACEBOOK:

The best way to report abusive content or spam on Facebook is by using the Report link that appears near the content itself. **https://www.facebook.com/help/www/181495968648557**

To report a post:
1. Click in the top right of the post
2. Click **Report post** or **Report photo**
3. Select the option that best describes the issue and follow the on-screen instructions

Report Something on Facebook
Please select the option that best describes what you'd like to report, then use the link or information provided to find the best way to report it. By choosing the correct option, you'll help us to review your report faster and more accurately.

What best describes the issue that you want to report?

- My account has been hacked
- I want to report something that shouldn't be on Facebook (e.g. photo, group, Page)
- Bullying or harassment
- Other abuse

Tools for Addressing Abuse on Facebook: **https://www.facebook.com/help/359033794168099/**

5 Things You Can Do
1. Send a message to the person responsible for posting
2. Unfriend the person to remove them from your friend list
3. Block the person from contacting you
4. Report the person if their behavior is abusive

Use privacy settings
We're sorry you're having a bad experience on Facebook, and we want to help. If you want to report something that goes against our Community Standards (example: nudity, hate speech, violence), use the Report link near the post, photo or comment to report it to us.

If you want to report something that goes against our Community Standards but you don't have an account or can't see the content (example: someone blocked you), you may need to ask a friend to help you. Remember that you should contact local law enforcement if you ever feel threatened by something you see on Facebook.

Staying Safe on Facebook: **https://www.facebook.com/safety/tools/safety**

Safety on Twitter: Our approach
https://about.twitter.com/en_gb/safety.html

Free expression is a human right. Everyone has a voice and the right to use it. On Twitter, you should feel safe expressing your unique point of view with every Tweet – and it's our job to make that happen.
But sometimes Tweets can cross a line and are abusive or threatening. To keep you safe, we build tools so you can control what you see and who you interact with; work with a community of online safety experts to fight abuse everywhere; and develop and enforce policies to prohibit abusive behaviour.

HOW TO REPORT ABUSE ON TWITTER:

https://support.twitter.com/forms/abusiveuser

Someone on Twitter is engaging in abusive or harassing behaviour.
Please fill out all the fields below so we can review your report.

What are you reporting?

- Harassment
- Specific violent threats involving physical safety or well-being
- Exposed private information or photo
- Someone on Twitter is posting spam
- Directs hate against a race, gender, religion, or orientation

These actions are...

- Directed at me (e.g. @mention, name, nickname or pseudonym)
- Directed at someone I legally represent (e.g. a client or my child)
- Directed at others (e.g. a friend or group)

Report details
What username is causing the issue? @
(e.g. @safety) Please provide specific Tweets as evidence of this issue.

Reported Tweet URL
Instructions on how to find the direct URL to a Tweet.
If what you are reporting appears outside of a Tweet (e.g. account bio, profile photo or header), please provide details in further description of problem.

Further description of problem
Please provide as much detail as possible surrounding your issue. For example, if you think the user has multiple accounts to directly @reply you, please list them above.

We are unable to accept attachments or screenshots related to your report.
Please only provide links to exact Tweets or Twitter accounts.

Tell us about yourself:
Your email
This is the email we'll use to contact you. Enter your current address.
Twitter username (optional) @
Signature
(Please electronically sign this notice by typing your full name)

How do I file a report that someone is abusive via Tweets or Direct Messages?
https://support.twitter.com/articles/20169998

Anyone can report abusive behaviour directly from a Tweet, profile, or Direct Message.

To report a Tweet:

4. Navigate to the Tweet you'd like to report on twitter.com or from the Twitter for iOS or Android app.
5. Click or tap the icon.
6. Select **Report**.
7. Select **It's abusive or harmful**.
8. Next, we'll ask you to provide more information about the issue you're reporting. We may also ask you to select additional Tweets from the account you're reporting so we have better context to evaluate your report.
9. Once you've submitted your report, we'll provide recommendations for additional actions you can take to improve your Twitter experience.

To report an account:

1. Go to the account profile and click or tap the gear icon (iOS), or tap the overflow icon (on twitter.com and Android).
2. Select **Report**.
3. Select **They're being abusive or harmful**.
4. Next, we'll ask you to provide additional information about the issue you're reporting. We may also ask you to select Tweets from that account so we have better context to evaluate your report.
5. Once you've submitted your report, we'll provide recommendations for additional actions you can take to improve your Twitter experience.

To report an individual message or conversation via twitter.com:

1. Click into the Direct Message conversation and find the message you'd like to report. (To report the entire conversation, click the **more** icon)
2. Hover over the message and click the **report** icon when it appears.
3. Select **Report @username**.
4. If you select **It's abusive or harmful**, we'll ask you to provide additional information about the issue you're reporting. We may also ask you to select additional messages from the account you're reporting so we have better context to evaluate your report.
5. Once you've submitted your report, we'll provide recommendations for additional actions you can take to improve your Twitter experience.

What should I do if I receive a violent threat?

You can report Tweets, profiles, or Direct Messages directly to us. Twitter may take action on the threatening Tweet, Direct Message, and/or the responsible account.

However, if someone has Tweeted or messaged a violent threat that you feel is credible or you fear for your own or someone else's physical safety, you may want to contact your local law enforcement agency. They can accurately assess the validity of the threat, investigate the source of the threat, and respond to concerns about physical safety. If contacted by law enforcement directly, we can work with them and provide the necessary information for their investigation of the threat.

For Tweet reports only: You can get your own copy of your report of a violent threat to share with law enforcement by clicking **Email report** on the **We have received your report** screen.

#defyhatenow social media #Peacejam "Unity can bring peace", Rhino Camp Uganda 2016

Reporting hateful content on YouTube:
https://support.google.com/youtube/answer/2801939

We encourage free speech and try to defend your right to express unpopular points of view, but we don't permit hate speech. Hate speech refers to content that promotes violence against or has the primary purpose of inciting hatred against individuals or groups based on certain attributes, such as: race or ethnic origin, religion, disability, gender, age, veteran status, sexual orientation/gender identity.

Keep in mind that not everything that's mean or insulting is hate speech. If you're upset by content that a specific person is posting, you may wish to consider blocking the user. However, if you feel that content violates our hate speech policy, report it to YouTube for review in one of the following ways:

Flag the video :
You may report hateful content that you think may violate our community guidelines by flagging it.

Mobile

1.Go to the video you'd like to report.

2.Tap More at the top of the video.

3.Tap **Report.**

4.Select the reason for flagging.

Computer

1.Log in to YouTube.

2.Below the player for the video you want to flag, click **More**.

3.In the drop-down menu, choose **Report**.

4.Select the reason for flagging that best fits the violation in the video.

5.Provide any additional details that may help the review team make their decision including timestamps or descriptions of the violation.

File an abuse report : If you have found multiple videos, comments, or a user's entire account that you wish to report, please visit our reporting tool, where you will be able to submit a more detailed report.
https://www.youtube.com/reportabuse

Safety and Abuse Reporting
What is the issue?

- Harassment and Cyber bullying
- Violent Threats
- Child Endangerment
- Hate Speech Against a Protected Group

Violent Threat
We want to ensure that YouTube is a safe place for our users while allowing for a vibrant community to flourish. While some content may be insulting or offensive, please note that we will only remove serious threats. Enter channel URL of the user you want to report:

Harmful or dangerous content: **https://support.google.com/youtube/answer/2801964**

While it might not seem fair to say you can't show something because of what viewers might do in response, we draw the line at content that intends to incite violence or encourage dangerous or illegal activities that have an inherent risk of serious physical harm or death. Videos that incite others to commit acts of violence are strictly prohibited from YouTube. If your video asks others to commit an act of violence or threatens people with serious acts of violence, it will be removed from the site.

+

INFORMATION ON HATE SPEECH

DW Hate Speech FAQ
http://www.dw.com/en/hate-speech-a-faq/a-19103744

UNESCO REPORT ON COUNTERING ONLINE HATE SPEECH
http://unesdoc.unesco.org/images/0023/002332/233231e.pdf

No Hate Speech Movement
"The campaign is against the expressions of hate speech online in all its forms, including those that most affect young people, such as forms of cyber-bullying and cyber-hate. The campaign is based upon human rights education, youth participation and media literacy. It aims at reducing hate speech and at combating racism and discrimination in their online expression."
http://www.nohatespeechmovement.org/survey

Hate Speech Explained, A Toolkit
Article 19, 2015 CC-BY-SA
http://www.article19.org/pages/en/hate-speech.html

PeaceTech Lab Social Media Lexicon of Hate Speech Terms
http://www.peacetechlab.org/hate-speech-in-south-sudan/

OSRX OPEN SITUATION ROOM

http://www.osrx.org/web/open-situation-room/ssd-hate speech

Project Summary, live visualizations from social media analysis of hate speech, news articles, and resources for organizations countering hate speech: South Sudan Hate Speech Data Portal.

Conflict Sensitive Resources

African Centre for the Constructive Resolution of Disputes (ACCORD)
http://www.accord.org.za/news/accord-trains-refugee-camp-leaders/

Youth Social Advocacy Team (YSAT) #defyhatenow Cultural Event, Rhino Camp Uganda