Using social media in CBP - Chapter 5

Moderation and Sensitive Content

Introduction

This chapter looks at how to use Social Media for conversations around sensitive issues that affect PoCs' rights. It explains how to moderate content online and how to use Social Media for constructive dialogue.

Forced displacement is often a politicized issue that affects how people of concern are perceived by host communities, and vice-versa. Sometimes these perceptions manifest as polarized conversations on Social Media, with the risk of adverse consequences for PoCs and their human rights.[83] Given UNHCR's role and visibility, the organization is often at the receiving end of anger and frustration. Sometimes these exchanges happen directly on our CBP Social Media accounts.

Here "sensitive content" means content that can harm PoCs, UNHCR or partners. Sensitive content is normally:[84]

- **Contextual:** What is sensitive in one context may not be in another;
- **Temporary:** Sensitivity will change over time, increasing or decreasing;
- **Relational:** Information may not be sensitive per se but can become sensitive if combined with other data, or if it is taken out of a dataset.

In this Chapter, we address "sensitive content" along with content that is "violent" or "offensive", including hate speech, derogatory remarks, insults, etc.

Engaging with PoCs on Social Media, particularly to moderate discussions, is difficult but very important. Content moderation is the act of applying guidelines to text, images and videos that appear on Social Media accounts or websites, often with a focus on user submissions.

Content moderation is used here to indicate:

- Actions like monitoring and identifying potentially harmful content;
- Assessing whether content complies with the site's guidelines, Community Rules, Code of Conduct and relevant UNHCR Policies and Guidance;
- Actions that involve UNHCR or partners in conversations that could potentially lead to harmful content or behavior. The intent is to influence the conversation positively;
- Actions to support peace building and reconciliation, and to counter xenophobia and racism.

83 Keep in mind that here we are not just talking about protection concerns (threats and risks) created by online activity, e.g., due to the data/information about individuals or the PoC community that is shared or disclosed. We also have in mind information used to influence the way PoCs and host communities perceive each other; as a way to increase discrimination against certain groups; or as a way to influence decisions made by vulnerable populations in need.

84 Adapted from Protection Information Management (PIM) Initiative, PIM Training Resource Pack, Module 5.1: PIM Sensitivities, http://pim.guide/uncategorized/pim-training-resource-pack

Moderating content is a delicate task, often involving additional processes that go beyond the work of vetting violent or extremist content and protecting PoC's personal information. Moderation also carries risks, including reputational and organizational risks that should be spelled out in your Risk Assessment.[85]

1 Types of Content Moderation

Important

Normally, communities and individuals develop mechanisms to cope with and respond to the protection problems they face. In many situations, they will already be dealing with the problem, although they may welcome extra support. However, there may be situations where community members do not recognize a practice as a protection risk or a violation of human rights, and there will be no community response or the response might be inadequate. This is often the case for online conversations, where there is a sense of perceived anonymity and/or detachment from consequences. When the community response does not meet international human rights standards, we should work with people to change it.

The content moderation method that makes most sense for you will depend on your Social Media goals and tools, and the operational context. Among the various types of content moderation, the most common are:[86]

I. Pre-Moderation

This involves all user submissions being placed in a queue for moderation before they are displayed on the site/Social Media page. Through pre-moderation, it is possible to keep all sensitive content off a site by checking every comment, image or video. However, if the CBP program requires immediacy and barrier-free engagement, this method can make it look as if the platform is being censored, which may erode trust and limit transparency. This method is best for Social Media accounts that need high levels of protection, such as those frequented by children;

This type of moderation is possible on Facebook, where you ask to review all posts and comments before they go public.

85 See more on Risk Assessment in Chapter 2.

86 Daniel Smith, The Essential Guide to Content Moderation, August 05, 2019

II. Post-Moderation

When user engagement is important but a comprehensive moderation program is still required, post-moderation is often a good choice. It allows users to publish their submissions immediately but also adds them to a queue to be reviewed. This allows for immediacy but enables moderators to monitor behavior. This can put pressure on a moderator, who has to approve every comment. Sensitive issues posted publicly must be detected and removed very quickly, which means you need a 24/7 moderation system;

III. Reactive Moderation

For a manageable program that relies on the community, reactive moderation is a possible solution. This type of moderation asks users to flag any content they find offensive or that breaches Community Guidelines. By involving PoCs in the process, reactive moderation directs attention to the content that most needs it. However, there is a risk that offensive or sensitive content will remain online for longer, which could damage the reputation of the account (and organization). For reactive moderation, there must be substantial investment in training and capacity-building for both users and managers;

IV. Supervisor Moderation

Similar to reactive moderation, supervisor moderation involves selecting a group of moderators from the online community. Also known as unilateral moderation, the system gives certain users the right to edit or delete submissions. If supervisors are selected carefully, this method can remove sensitive content promptly and works well as the community grows. However, it is prone to fail if moderators miss offensive text, images or video;

This type of moderation is the one used by the UNHCR Lebanon office in managing their Facebook Community Page.

V. Commercial Content Moderation (CCM)

CCM involves outsourcing moderation to specialists, which for UNHCR could mean local organizations or NGOs. They would be tasked with ensuring that content abides by Community Guidelines, user agreements and the legal frameworks for the particular site and country. Since the work is done by specialists, a good standard of moderation is usually guaranteed. This way you can pass difficult or controversial conversations to trained experts;

This method is best suited to conflict situations or conversations that are highly polarized. See here an example of outsourced moderation.

VI. Distributed Moderation

As one of the most hands-off moderation systems, distributed moderation places a lot of trust and control with the community. It usually involves allowing users to rate or vote on submissions, flagging content which goes against guidelines. It often takes place under the guidance of experienced moderators and can work well if a Social Media page/group has a large, active community;

VII. Automated Moderation

Automated moderation is increasingly popular. As the name suggests, it uses various tools, including AI or artificial machine learning,[87] to filter, flag and reject user submissions. These tools can range from simple filters, which search for banned words or block certain IP addresses, to machine learning algorithms, which detect inappropriate content in images and video. If you decide to use these tools,[88] combine them with some kind of human moderation. Automated moderation can be also helpful if there are different languages on your site, as algorithms are getting better at working with multiple languages.

See an example of automated moderation, paired with a human component, on this WHO project to provide information about COVID-19.

How to create your own moderation strategy

No matter what type of moderation you choose, a moderation strategy for your Social Media content should always include the following elements, which are explained below:

a) Criteria to identify or flag content;
b) Strategies for different types of content identified;
c) Monitoring of the process to decide adjustments.

a) Criteria to identify or flag content

Having a clear matrix to indicate which content requires moderation not only helps to organize work internally but also builds trust in your transparency, if you share with your audience on Social Media. Normally content is flagged if it falls under these categories:

- Content that violates human rights, rule of law or code of conduct/community rules. This includes hate speech, offensive language and content the platform considers sensitive (e.g. depictions of blood or violence);

87 Find more information on Cambridge Consultants, Use of AI in online content moderation, report produced on behalf of OfCom, 2019

88 AI in general presents several issues that may be of concern for data privacy and security, and that may reinforce biases and exclusion. For this reason it is always best to involve ICT, UNHCR Innovation and DES in decisions about using these tools.

- Content that, while not containing anything violent or sensitive, could affect peoples' decision-making or actions. This is the most difficult to moderate because it may not be obviously dangerous. Most of the time this content boils down to misinformation, disinformation or rumors but it also includes Personal Data and information about protection incidents and locations, etc.

Resource

See Chapter 6 for more on how to handle rumors and misinformation on Social Media.

Within these categories you can specify what content you consider unsuitable and the types of conversations that would promote your objectives. The best way to do this is to engage the community itself in drafting a "Code of Conduct" or "House Rules" for the Social Media account/ page. This develops a sense of ownership from the outset.

b) Strategies for different types of content

Defining what content needs to be flagged is of course a participatory process. It directly affects our ability to raise the community's awareness about human rights, particularly the rights of women and children.
When deciding your strategy to moderate content, try whenever possible to:

- Use workshops and discussions to analyze the community's rights practices online. Which rights are being respected and by whom? It can be helpful to compare online human rights standards with community values and identify areas where they coincide. Discussion points can include:[89]
  a) whether all people can exercise their rights;
  b) if not, why some people are excluded;
  c) which rights are not being respected and why;
  d) who is a rights holder and who has a duty.

  This can lead to discussions about what the community needs to do to improve PoCs' enjoyment of their rights on Social Media. You can agree on what constitutes a protection risk online and how respect for individual rights on Social Media should inform any online protection response;

89 Here we are talking about using Social Media to engage the community on human rights issues in their context; also to discuss 'online rights' and issues such as privacy and confidentiality.

- Discuss possible responses to moderated content with the community concerned. Facilitate discussions with community members about how offensive or sensitive content can negatively affect individuals and have an impact on family and/or community;
- Find the root of a negative or harmful practice and ask why it is considered important or valuable. Identify possible opportunities for consciousness-raising. Which individuals or groups might be willing to work for change?
- Ensure that people have understood which online practices are unacceptable and why, and see to it that UNHCR and its partners do not support such practices;
- In cases where the community does not recognize the harm a practice could do to an individual, UNHCR should intervene directly. This requires careful consideration because a proper response must go beyond immediate safety and/or the restitution of rights and avoid negative consequences both for those affected and those intervening.

c) Monitoring content moderation to inform adjustments

It is important to monitor the frequency and topics of sensitive content being posted on our pages/groups. Apart from doing an AGD analysis of the participants, it is useful to look at how their conversations relate to the situation on the ground. How do users behaving in a certain way respond to UNHCR policies and communication guidelines?

Monitoring the moderated content can also clarify the extent to which additional resources may be needed or moderation strategies must be adjusted.

Often sensitive topics surface online because of:

- the operational context on the ground – this could be a new influx of refugees from a neighboring country or an event that hardens public opinion against IDPs;
- the ability of UNHCR and partners to educate and involve communities in the process;
- the online rights landscape in the country, where regulation of online behavior and enforcement may vary (see Chapter 1).

2. Moderating Angry, Abusive or Inappropriate Comments

Going through conflict, abuse and forced displacement can make persons of concern very emotional and at the same time unable to put their feelings into words. Accountability means taking them seriously and listening to what they have to say, regardless of whether we agree or disagree with, and like or dislike, their views.

Communication can be difficult if anger, insecurity or fatigue, plus, in some cases, having to speak in a foreign language, impede PoC's ability to express themselves. Under pressure, distressed people may say things that seem offensive or inappropriate without meaning to be aggressive.[90]

When dealing with such angry/abusive comments or posts, consider the following:

If this is the first time you have interacted with the person, or the first time the person has interacted in this way, always give him/her the chance to understand why their message comes across as offensive and to engage differently next time. Reach out to and engage (privately, if possible) with the person to understand if their offensive language may have come out of frustration or simply be a way to attract attention;

If the inappropriate content is posted publicly by an influencer or someone known and recognized as an authority on the matter, don't remain silent but immediately engage publicly and directly. (If there are possible consequences for UNHCR's reputation and brand, tackle this with support from External Relations);

Always put the person talking to you at the center of the interaction and ensure that they feel heard, valued, understood and respected. Paraphrase what they have said in constructive, positive language to show you have listened and understood correctly and that it matters to you. Acknowledge different viewpoints and reconcile potential misunderstandings. Try to answer questions as concretely as possible. Observe how the conversation evolves and if your de-escalation techniques are working. Sometimes people just want to vent and allowing them to do that may be all they need.

90 See more about this on UNHCR, Community-Based Protection in Action - Effective & Respectful Communication in Forced Displacement, 2016

Tips

When dealing with angry remarks, remember:

- Summarizing angry remarks (removing inappropriate language, if there is any) may clarify and reflect their feelings, while de-escalating conflict;
- Restating thoughts calmly shows you understand their feelings;
- Being patient allows for the full expression of anger;
- Accepting the other's emotions creates trust;
- The anger is likely to be aimed at the situation and not at you personally;
- Putting yourself in the other person's shoes helps you gain perspective on their situation;
- Asking questions purposefully clears up misunderstandings and makes the person feel understood.

From UNHCR, Community-Based Protection in Action - Effective & Respectful Communication in Forced Displacement, 2016

It is a different matter if the offensive messages constitute a pattern, meaning you could be dealing with a troll. Then the best approach is to remove content, always posting a message explaining why you have done so.

Comments or messages that violate the rule of law or incite to violence:

- If someone posts hateful messages, or threatens you, your colleagues or PoCs, delete the comments;
- If needed, print the comments or make screenshots and inform the UNHCR security team;
- All Social Media platforms have mechanisms through which you can report abusive behavior.


© UNHCR/Jaime Giménez Sánchez de la Blanca

3. Moderating Polarizing Conversations

Over the last decade, researchers working to understand the impacts of emerging information and communications technology (ICTs) posit that political groupings have become siloed, polarizing public discourse. Moving people from passively accepting something that escalates conflict to trying to achieve dialogue in their society is a huge challenge. But numerous initiatives that leverage ICTs to encourage constructive online dialogue have emerged around the world.[91]

Resource

Find information on how to deal with tension and conflict between groups of People of Concern (section 1.3) and tension and conflict between People of Concern and host communities (section 1.4) in the "UNHCR Manual on Security of People of concern"

No matter what your project, and what Social Media you use, you will probably have to moderate polarizing or political conversations on the rights of Persons of Concern.

UNHCR's role in managing and engaging with PoCs on digital platforms is to make sure interactions take the form of a dialogue. We want participants to read and respond to each other's comments, asking questions for mutual understanding rather than to prove a point. In this way they can build on each other's ideas.

Different strategies work best when combined together:

- **Leading by example:** UNHCR leads by example in the way it speaks and engages online. Entering polarizing conversations to try to change the tone and narratives that accompany conflicts has proved to be very effective;
- **Online education activities:** UNHCR has a leading role in educating PoCs to engage constructively online. This starts with clear and well explained Community Rules for our Social Media presence. But active engagement can also take the form of polls, online quizzes or dedicated educational campaigns focusing on human rights;

91 Build Up, Building The Commons, 2018

- **Offline activities:** This encompasses all the usual work UNHCR does in a country on human rights, xenophobia and protection, from dealing with governments and law makers to helping persons of concern and grassroots organizations on the ground.

Resources

In May 2020, Over Zero developed a guide called "Counteracting Dangerous Narratives in the Time of COVID-19" to offer insights into preventing worsening division and identity-based violence. Access the full guide here and the summary here.

Generally speaking, the following online behaviors support dialogue:

- **Suspension of judgment while listening and speaking:** When we suspend judgment and listen, we open the door to understanding. When we speak without judgment, we encourage others to listen to us;[92]
- **Respect for differences:** Everyone has an essential contribution to make and is to be honored for the perspective that only they can bring;[93]
- **Role and status suspension:** All participants and their contributions are essential to the whole view. No one perspective is more important than another. Dialogue is about 'power with', not 'power over';[94]
- **Balancing inquiry and advocacy:** In dialogue, we inquire to understand another's perspectives and offer our own for consideration. The intention is to bring out assumptions and relationships and gain new insights;[95]
- **Focus on learning:** We aim to learn from each other and expand our understanding, not compete to see who has the "best" view. When we are focused on learning, we tend to ask more questions and try new things. We are willing to open up to see what is working for us and what we might want to change. We want to hear from all parties to get the advantage of differing perspectives.[96]

Whether you are doing a Social Media project that tackles online polarization or dealing with sporadic interventions on your SM account, engaging conflicted communities to discuss protection issues on Social Media demands a lot of skills and requires all the actors to work together. Within UNHCR, you can always reach out to DER, DIP and DRS.

92 Center for Whole Communities, A Brief Orientation to Dialogue, 2006
93 Alison Jones, Speaking Together: Applying the Principles and Practice of Dialogue, 1996
94 Portland University, Capstone Project, The Difference between Dialogue and Debate, unknown
95 Bruce Mohst, Choose your Communication Tool — Debate or Dialogue?, 2015
96 Gary Hall, The Importance of Questioning, 2016

Resourcing

Online content moderators need the same skills as traditional meeting moderators, plus familiarity with the SM platform used. Consider the following when determining the resources you will need for moderation on Social media:

- **Know the landscape:** Make sure you have used your Situation Analysis[97] to understand the existing landscape when it comes to digital rights and how people interact online with each other and with UNHCR. Knowing what to expect is the first step to designing a tailored response strategy;

- **Go hyper-local when you can:** Moderation and conversations about sensitive issues online are strongly affected by the language used and by cultural perceptions. Online moderators should always be from the same culture and speak the same language as the community. If you can, and after considering risks of bias and polarization, and possible safeguarding issues, engage them directly from among the PoCs you work with;

- **Always engage with Public Information/External Relations and colleagues in the operation:** Online conversations about sensitive topics and human rights can affect the way UNHCR is perceived in the country, and by PoCs. Since they manage the operation's Social Media accounts, External Relations have the experience to help you deal with sensitive conversations online;

- T**ake every chance to engage your community:** If you find yourself removing content for any reason, explain yourself and engage your community to discuss why the content had to be removed. Use that opportunity to suggest why certain tones/words are unacceptable;

- **Be honest:** If there are topics you cannot or do not want to discuss, make it clear why, e.g. via House Rules. The more you clarify the boundaries of your engagement with people, the less you will have to moderate it;

- **Always connect with offline work:** UNHCR is often involved in projects that put peace-building, peaceful coexistence and integration at the heart of efforts to protect PoC's rights. These projects take the form of guided dialogues and other community activities on the ground. Make sure you always connect with these activities to have a coordinated approach to community engagement.

97 See Chapter 1 for more information on Situation Analysis

Do's	Don'ts
DO get together with the community to identify what content is sensitive. Consider cultural values when it comes to online rights and conversations.	**DO NOT** assume you know the intentions of your interlocutors, even if you think they are being unfair or might even be trolling.
DO respond and stay calm, no matter how angry or upset the other person is. Always provide opportunities for a constructive dialogue.	**DO NOT** delete messages without explaining why to the rest of the community.
DO make maximum use of local partners and PoCs to manage Social Media projects that address peaceful co-existence, reconciliation and integration.	**DO NOT** set up a Social Media account without having an online moderation strategy planned.
DO use the right advocates within the community, who are normally far more effective than external figures, such as UNHCR staff.	**DO NOT** expose your advocates or staff to security risks and always make sure you update your Risk Assessment.
DO engage men and boys and invite them to online discussions on issues such as masculinity, fatherhood, gender equality, reproductive health, HIV/AIDS and sexual and gender-based violence. This fosters the equality and empowerment of women.	**DO NOT** shut down a conversation just because UNHCR lacks an official position on a subject (or prefers not to make it public). You may not be speaking for UNHCR but you can still reflect a rights-based approach.
DO work with the community, ensuring all vulnerable groups are included and represented. Make the community aware of their rights and obligations through accessible and appropriate content.	**DO NOT** assume that everyone knows what sensitive content is, and what constitutes hate speech.

Check List

- [ ] Have you discussed and agreed with all stakeholders, including communities, what needs to be moderated on Social Media, and why?
- [ ] Have you identified local actors, e.g. Civil Society Organizations (CSOs), already moderating sensitive issues on Social Media from whom you could learn?
- [ ] Do you have a grasp of the resources, challenges, regulations and law around hate speech, incitement to violence or any other limitations to freedom of speech?
- [ ] Have you identified what is 'sensitive' content in your operational context?
- [ ] Have you coordinated and discussed your moderation requirements with external relations?
- [ ] When discussing your moderation strategy, did you involve communities and stakeholders to ensure diverse representation?
- [ ] Have you used an AGD lens to look at risks, possible victims of online rights violations and possible solutions?
- [ ] Do you know the terms of reference or community guidelines that apply to the Social Media platform you want to use?
- [ ] Have you assessed the existing skills and capacity to moderate online conversations within UNHCR and with partners?
- [ ] Have you created and shared widely your community guidelines (House Rules) for content moderation on your Social Media channel?
- [ ] Have you worked out when and how often you will train staff on managing sensitive content and abusive conversations? What skills might they need to develop, strengthen or refresh?

© UNHCR/Jaime Giménez Sánchez de la Blanca

Case Studies

Build Up, The Commons, a Pilot Methodology for Addressing Polarization Online, 2018

UNDP, Analyzing Refugee-Host Community Narratives on SOCIAL MEDIA in Lebanon, 2019

UNHCR Innovation, Teaching a 'Robot' to Detect Xenophobia Online,2018

UNHCR, Protection from Xenophobia, An Evaluation of UNHCR's Regional Office for Southern Africa's Xenophobia Related Programs, 2015