ONLINE COMMUNITY MODERATION

THIS IS AN RNW MEDIA
KNOWLEDGE BRIEF

August 2020

digital communities for social change

TABLE OF CONTENTS

1. INTRODUCTION

This knowledge brief showcases RNW Media's experience with online community moderation. It explains what community moderation means, why we believe it is an essential component of our Citizens' Voice programme, and the strategies we have developed together with our moderators to guide them in the work they do and the decisions they need to make. Also, we present results from research that we conducted on three of our communities to test the effectiveness of our approach. Lastly, our moderators working in Burundi, DRC and Mali share their experience and insights, showing how moderation goes in practice. They talk about the secret of being a good moderator and share the challenges they experience in their daily work.

Citizens' Voice

Citizens' Voice is active in Burundi, China, DRC, Egypt, Libya, Mali and Yemen and focuses on social cohesion and inclusive governance. Applying a user-centred approach and persuasive storytelling, Citizens' Voice creates and maintains digital platforms offering safe spaces where young women and men from across political, ethnic, racial, regional or religious divides can come together in a way which is often impossible in the offline space. In-country teams create multi-media content relevant to the local context to attract and engage young people via a variety of digital channels. Through moderated discussions, they encourage disparate groups to voice their opinions on issues of common interest both online and offline. To achieve impact, Citizens' Voice aims to engage the broadest possible spectrum of stakeholders at all levels. By building inclusive communities where everyone feels safe in otherwise fragmented societies, an alternative civic space is created and can stimulate the move from polarised discussion to constructive debate and dialogue.

Online Community Moderation

Online media are increasingly facing challenges regarding inappropriate, negative and/or low-quality comments in their discussion sections. To tackle this issue, some online media decided to disable the comment function below articles altogether, other media do this for (sensitive) articles that are likely to elicit unconstructive comments. In order to deal with the challenges, RNW Media's Citizens' Voice programme created a moderation strategy. This strategy gives guidance to moderators to facilitate constructive and inclusive discussions on the pages. The moderation strategy is an essential component of the programme as one of the key objectives of Citizens' Voice is to contribute to social cohesion among young people in fragmented societies. Our platforms aim to be (online) spaces where young women and men from across political, ethnic, racial, regional or religious divides can respectfully discuss all kinds of topics. Citizens' Voice believes that when young people recognise that there are multiple viewpoints, this will contribute to more openness and acceptance towards peers who hold different viewpoints, increasing social cohesion among youth. Also, it will lead to young people challenging the restrictive norms and prejudices that exist in their country. Since content on our platforms regularly addresses sensitive topics, including discussion around existing socio-cultural norms, a sophisticated moderation strategy is key in facilitating constructive discussion and inclusive dialogue among young people in our digital communities.

Defining online moderation

Ziegele & Jost (2016)[1] define online moderation as: *"Any kind of institutional engagement aimed at the organisation or regulation of the processes or contents of online discussion".*

They explain that there are (at least) three types of online moderation:

1. ***Collaborative moderation:*** members of an online community rate or flag each other's comments and inappropriate contributions will either receive less attention, are automatically removed after a certain number of flags or will be sent for review.

2. ***Content moderation:*** moderator(s) intervene in the online discussion by deleting (or hiding) inappropriate contributions. Also, moderators could decide, prior to publication of a comment, whether to allow the comment to become visible to other members of the community, something which is called pre-moderation.

3. ***Interactive moderation*** moderator(s) engage with commenters by clarifying user questions, appreciating user's contributions, challenging extreme opinions and keeping discussions on topic.

[1] *Ziegele, M., & Jost, P. B. (2016). Not funny? The effects of factual versus sarcastic journalistic responses to uncivil user comments. Communication research, 0093650216671854.*

2. CITIZENS' VOICE MODERATION STRATEGY

Our approach

For RNW Media, moderation goes deeper than removing extremist content, it's about looking at the conversations taking place in our communities and generating dialogue. The moderation strategy is based on choices about how we want our online communities to interact and on encouraging positive behaviours that result in constructive discourse and inclusive dialogue. This enables safe spaces where young people can express their views, needs and opinions and engage in respectful discussion about topics that matter to them. Citizens' Voice mainly applies an interactive moderation approach to its communities. Our moderators also fulfil the role of regulator by trying to move away from polarisation and by publicly referring disruptive commenters to the community guidelines. Apart from removing spam or commercial posts, moderators are cautious when moderating content in order to respect the principle of freedom of speech. However, to ensure communities remain a safe space, the moderators will hide comments that are violent, abusive or discriminatory, in combination with a direct message to the commenter. Special attention is paid to creating and maintaining a safe environment for women to engage in the communities.

Trained moderators

Each Citizens' Voice project has trained moderators following all the discussions taking place on the different digital media channels. The matrix below provides guidance to moderators on how to respond to different types of comments. The horizontal axis shows the type of comment from platform users, the vertical axis shows how the moderator could respond.

Supportive / Constructive	Inquisitive	Negative / Unconstructive	Antagonistic	Abusive / Offensive	MODERATOR RESPONSE
X					Like / React
X					Positive Affirmation
	X				Answer Question
	X	X			Inform / Clarify
X		X	X		Ignore
		X	X		Challenge
		X	X		Enforce Community Guidelines (Publicly)
			X	X	Hide Comment
			X	X	Hide (with DM to user)
				X	Hide & Block

Objectives, Strategy and Tactics

Objective	Strategy	Tactics[2]
Create a safe and inclusive environment	Young people feel able to join the conversation and feel heard. They see and understand the benefits of respectful dialogue.	➢ Reinforce (through acknowledgment) constructive conversation and behaviour of users. ➢ Hide violent or inflammatory comments ➢ Ignore unconstructive comments so they don't get extra visibility. ➢ Challenge, hide, or ignore antagonistic comments (depending on the infringement/moderator's judgement) ➢ Warn users not abiding by the community guidelines and, ultimately, block them.
Provide additional information	Sometimes users have questions after reading the content and moderators can answer these or refer users to specific website articles.	➢ Answer users' questions ➢ Provide additional information (when requested/required) ➢ Validate users' opinions ➢ Settle disputes and tackle misinformation with facts or research
Stimulate constructive and respectful dialogue	It's important to allow space for different perspectives. We encourage users to reflect on their own standpoint and become open to different perspectives. It's the moderator's job to get that discussion going and encourage a robust, critical conversation.	➢ Encourage users to share their opinions ➢ Prompt people (of opposing viewpoints) to engage with each other ➢ Encourage users to read the content and/or refer them to specific related content ➢ Demonstrate how to respect the viewpoints of others ➢ Bring users back on topic if discussions stray away from the original topic ➢ Support and create space for different opinions to be expressed
Diffuse polarisation	We do not shy away from polarisation on core topics but look to manage the conversation and move them from monologues to a constructive dialogue. Polarisation creates an atmosphere where users do not feel able to participate, we aim to diffuse such situations.	➢ *Change the target audience.* Pushers see those with opposing views as enemies and target the middle ground to intensify polarisation. So, target the middle ground for depolarisation; This could take the form of ignoring extreme, polarised positions and looking and highlighting the opinions of the middle. ➢ *Change the topic.* Move away from the identity construct chosen by the pushers and start a conversation on the common concerns and interest of those in the middle ground; Apply the aspirational approach and get the conversation back on the issues that all young people are experiencing. ➢ *Change position.* Don't act above the parties, in between the poles, but move towards the middle ground; Stop trying to build bridges (positioning above the poles) but rather to a position in the middle (connected and mediating). ➢ *Change the tone*; Use mediating speech and try to engage and connect with the diverse middle ground. Moderators should not moralise or lay blame but should focus on the development of mediating speech and behaviour
Gauge the temperature and keep users and staff safe	We adjust our content and social media strategy when the situation in our target country is particularly tense to guarantee the safety of our staff and users.	Country specific but could include; ➢ Manually approve all comments on the page (not a long-term, nor particularly scalable solution) ➢ Re-evaluate the topics of content we distribute (could be avoiding divisive content) ➢ When the situation in the country is too 'hot' we should take steps to guarantee the safety of our staff and our users.

[2] Based on Brandsma's model on Polarisation. Brandsma, B. (2017). *Polarisation: Understanding the Dynamics of Us Versus Them*. BB in Media.

3. MODERATION RESEARCH

Method & Research Question

***Research question:* "How does online community moderation affect user engagement?"**

In order to test the effectiveness of the Citizens' Voice moderation strategy and to understand how moderators influence an online discussion, we conducted nine experiments (A/B tests) on the Facebook pages of Yaga (Burundi), Habari RDC (Democratic Republic of Congo), and Benbere (Mali), between December 2018 and December 2019. Five experiments were conducted on Yaga, three on Habari RDC, and one on Benbere. This research[3] provided us with interesting insights.

A/B testing

A/B testing is a form of randomised controlled trial (RCT) in which you compare two groups against each other. Both groups are exposed to a different version of the experiment. In our case, a moderated post, or an unmoderated post. All other factors are kept the same. Differences in the outcomes (discussions) between group A and B could explain the effect of a certain intervention.

Findings

Moderation elicits more thoughtful comments

Comments in the moderated version of a post turned out to be of higher quality. In the moderated version the commenters were twice as likely to comment with a *Thoughtful* comment as commenters in the unmoderated version (statistically significant). Thoughtful comments are user comments related to the article or another user's comment. They provide (new) information or an opinion and consist of more than one or a few words.

Moderation generates interaction

Commenters in the moderated version are nearly four times as likely to make a comment replying to another comment (interaction), instead of directly commenting on the generic post. Both findings are statistically significant. Although we observed a higher number of total user comments, more words used

[3] *Interested to read the full study, please contact pmel@rnw.org*

per comment, and more comments made by women in the moderated versions, these results were not statistically significant.

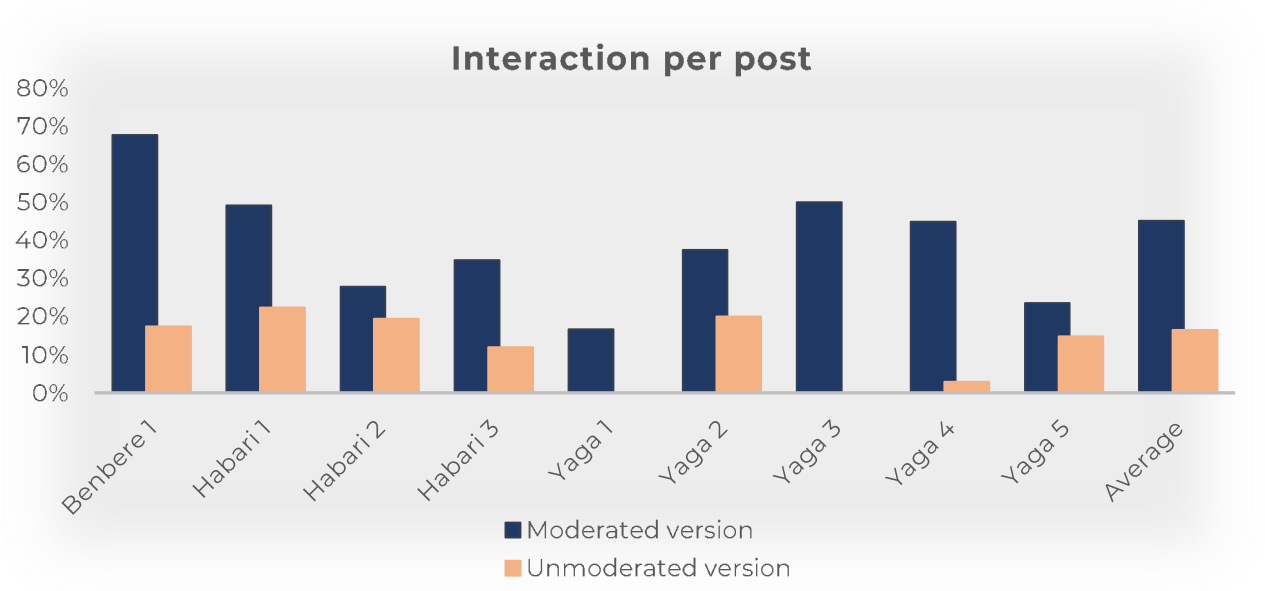

Figure 1. Y-axis shows the percentage of comments in the discussion that are directly addressing another user's comment, rather than the generic post.

Be aware of gender differences

Our data showed that men are more likely to respond with a *thoughtful* comment compared to women. On the other hand, women are more likely to respond with a *feedback* comment, a one- or two-word response indicating (dis)agreement/feeling with the post or another comment. This might suggest that women are more hesitant to express themselves online and prefer to amplify someone else's comment, or the post. Therefore, a moderator should keep in mind these gender differences when moderating a discussion.

Monitor extra closely once a discussion is "on"

The moderated discussions generated (contrary to our expectation) a slightly higher proportion of unconstructive comments compared to the unmoderated versions. One possible explanation is that moderating discussions contributes to more interaction in the discussion, leading to a more thorough exchange of viewpoints, and therefore also to more unconstructive comments. Our analysis revealed that sensitive posts are particularly likely to elicit unconstructive comments.

The tone is set at the beginning of a discussion

We found signs of an "Imitation effect" - that commenters are likely to copy each other's behaviour. Early comments made by users seem to influence how subsequent commenters respond. For example, sometimes comments are (literally) re-used by later commenters. Also, if the first few comments are emoticons, later comments seem to become more likely to also consist of an emoticon. Destructive comments at the beginning of a post could have a detrimental effect on the rest of the discussion, while constructive comments could positively impact the rest of the discussion. Steering a discussion towards a positive and constructive tone should be done early in the discussion.

4. MODERATORS' INSIGHTS

To better understand the work of community moderators, we talked with the moderators of Yaga, Habari and Benbere. They shared their insights as moderators of online communities. What is the secret of being a good moderator? What challenges do they face in their work? How to respond in specific situations?

Yaga (Burundi)

Consider the different personalities

"The most important role for a moderator is to bring different groups of people together and respect the different opinions they have. It's essential to consider the different personalities and backgrounds of our users", Elodie, moderator at Yaga, describes the key role of a moderator.

Keep users safe

Bella, moderator at Yaga (Urukundo), stresses the importance of keeping the users safe, *"We work a lot with taboo topics focussing on SRHR. Therefore, we need to be aware of the language that our users are using online. If our users use the "wrong" language* [in the eyes of the government], *it could have negative consequences for them and could even lead to arrests. There is an important role for us as moderators to closely follow the conversations and to propose other words they could safely use to still convey their message. In some cases, we must hide a comment of a user to protect their personal safety"*.

Take care of yourself

For Elodie the secret of a good moderator is to be patient, comprehensive and to take care of yourself. Always remember that you respond as an organisation and don't take threats personally. Bella's explains that in some cases she is tempted to respond from her own standpoint, *"For example, when a user thinks that women are inferior, I am tempted to answer as Bella (an individual), but I need to react as the organisation. I am not allowed to push for an argument in order to abruptly change the user's opinion"*.

Choose your words carefully

When a user posts hate speech on the Yaga page, the moderators try to make this person realise that they are offending people. However, it is not always easy to come up with the right words in Bella's experience, *"Especially because your words could be interpreted differently by the user"*. If a comment is clearly provoking other users, the moderators hide this comment and send a private message explaining that certain comments are not allowed on this platform, *"Yaga is a place to discuss, not a place to insult"*.

Be gender sensitive

Bella and Elodie experience a difference in commenting behaviour between men and women. For example, they see that in general the comments of women are shorter and that topics related to politics elicit very few comments from women, something that they believe is caused by the culture in Burundi in which women are not encouraged to speak in public. *"Last week we posted an article about handling intimacy. Most comments were made by men. It makes me happy when I see a woman responding and I try to make her at ease, for example by responding with an emoji"*. Although there are sometimes few comments made by women because they are more hesitant to respond publicly, the team knows that there are a lot of silent women readers on the Yaga pages. *"We get private feedback from users in our inbox in which they tell us that they appreciate the topics"*.

Benbere (Mali)

Neutralise trolls by argumentation

For Mohammed, moderator at Benbere, the most rewarding aspect of being a moderator is when he succeeds in persuading the most radicalised followers and trolls to see reason. Stimulating them to make constructive comments and accept the views of others. At Benbere moderators consider someone a troll when they respond to everything in an extremely negative way. *"In the beginning we had a lot of trolls, but by using argumentation they have moved away"*. Blocking a troll is not the solution Mohammed believes, *"we only block users that share advertisements"*.

> *"In the beginning we had a lot of trolls, but by using argumentation they have moved away"*

How to respond to accusations

It happens sometimes that users accuse Benbere of imposing Western values on them, values that are, in their eyes, not compatible with Malian culture. Mohamed gives the example of a discussion that took place around the topic of having many children. *"Many people in rural areas believe that having a lot of children is good, as they will take care of the parents when they are old, in that discussion we got accused of discouraging Malians from having many children, promoted by a Western agenda. We explained that having fewer children is good for other reasons. By having these discussions, the trolls move away"*. Another example he gives relates to a discussion they created around female genital mutilation. One of the comments Benbere received after they posted an article about a Malian woman who wrote a letter to her (lost) clitoris reads as follows: *"Be careful, they are descendants of SATAN on this planet, I hope Benbere will leave!!!"* Instead of hiding such a comment, the moderator would respond to this user. *"We started a discussion and asked them what the West would win by promoting that women are not circumcised?"* The moderators would explain that it is beneficial to women, since many women experience life-long pain as a consequence of FGM.

Tackle incorrect information

Another challenge Mohammed experienced is when users are spreading incorrect information. For example, regarding the existence of slavery in the Kayes region, *"Some people deny it, we respond to them by sending links to websites and information on this topic"*.

"Have love and passion for Moderation"

The secret of being a good moderator is to stay cool and try to put yourself in the shoes of the other. *"Each person is different and has a different understanding. My advice is to have love and passion for moderation because moderation means helping to create a community where young people feel safe"*.

Habari (DRC)

The three main roles of a moderator

The moderators of Habari distinguish three important roles for a moderator:
1. The social role – *"be a role model in discussions by using an appropriate tone. For example, appreciating positive messages through likes or by thanking users for their contribution".*
2. The expert role – *"be a guide for the users who ask us questions. We do this by providing links, resources and giving them access to reliable and trustable sources. This could help to support the opinion of a user or to debunk a negative opinion".*
3. The organisers role: *"Set the rules of do's and don'ts, hide offensive comments, keep a discussion going by asking questions, bring users back on topic and protect users who face personal attacks".*

Temper extreme standpoints

The moderators of Habari and Amour Afrique focus specifically on the leaders/pushers in the discussion when polarisation takes place: *"It is through them that it is possible to calm down the situation and temper the extreme standpoints that are taken in the discussion".* Topics related to gender, homosexuality, (some aspects of) politics and religion are the topics that are most likely to generate polarised comments in the moderators' experience. *"Regarding gender topics, we take a women's perspective in order to put women at ease and give them space to unleash their feelings".*

Moderation requires a lot of self-control

The moderators are sometimes put in a difficult position. Although they respect the ideas and beliefs of users, a specific comment could hurt them just as much as any other user. They explain: *"It is difficult when a situation arises that is linked to the personal situation of the moderator. How should a female moderator react when a misogynous man insults women? Or when it is related to a painful story about their own tribe? That is where self-control and empathy play an important role. Putting yourself in the mind of the offending users, understanding how they perceive that situation and from there moderate the discussion. There is a great deal of psychology in that process".* The moderators explain that they get some of their inspiration from the *"Socratic method"*, a form of dialogue based on asking and answering questions that stimulates critical thinking and brings ideas and underlying presuppositions to the forefront. A second theory that helps the moderators is *"Questiology"* from Frédéric Falisse, a theory that considers the art of asking good questions.

There is no universal user

The moderators of Habari RDC and Amour Afrique explain that they take a different approach to the different users. *"With more than 250 tribes (all in one single country), perceptions and ways of living are different. The meaning and interpretation of a specific word could be different in the East compared to the West and vice-versa".* Therefore, it is essential for the moderator to understand the local context in which discussions take place.

> *"With more than 250 tribes, perceptions and ways of living are different. The meaning and interpretation of a specific word could be different in the East compared to the West and vice-versa"*

digital
communities
for social
change

www.rnw.org

For any additional information or questions, please contact:

PMEL@rnw.org / Info@rnw.org