toolkit

minority
rights
group
international

CREID
Coalition for Religious Equality
and Inclusive Development

The Hate Speech Crisis: Ways to start fixing it

A toolkit for civil society organizations and activists

Claire Thomas, Mihaela Cojocaru and Noah Rosenberg

Cover image: A person uses a hand to block out hate speech while focusing on countering hateful content online with positive messaging, March 2022. *Credit: Infinite Lux*

Lead editor: Claire Thomas
Design: Mihaela Cojocaru
Editorial support: Noah Rosenberg

Acknowledgements

Minority Rights Group International (MRG) gratefully acknowledges the support of all organizations and individuals who contributed to this publication, including:

Arsalan Ashraf, Bytes for All
Cecilia Bisogni, Minority Rights Group International
Glenn Payot, Minority Rights Group International
Haroon Baloch, Bytes for All
Htet Swe, Peace Point-Myanmar
Iqbal Haider Butt, Bargad
Noah Rosenberg, Minority Rights Group International
Pshtiwan Farag, Independent Media Organisation in Kurdistan
Salam Omer, KirkukNow
Shahzad Ahmad, Bytes for All
Sharoon Sharif, Catholic Commission for Justice and Peace

For further information please contact MRG. A CIP catalogue record of this publication is available from the British Library.

ISBN 978-1-912938-61-2

Published: May 2022

Minority Rights Group International
54 Commercial Street
London E1 6LT
United Kingdom

Tel: +44 (0)20 7422 4200

Email: minority.rights@minorityrights.org
Website: www.minorityrights.org

Minority Rights Group International

Minority Rights Group International (MRG) is a non-governmental organization (NGO) working to secure the rights of ethnic, religious and linguistic minorities and indigenous peoples worldwide, and to promote cooperation and understanding between communities. Our activities are focused on international advocacy, training, publishing and outreach. We are guided by the needs expressed by our worldwide partner network of organizations, which represent minorities and indigenous peoples.

MRG is registered as a charity and a company limited by guarantee under English law: registered charity no. 282305, limited company no. 1544957.

CREID
Coalition for Religious Equality and Inclusive Development

UKaid
from the British people

Minority Rights Group International (MRG) gratefully acknowledges the support of Mariz Tadros, Sue Schirmer and the Coalition for Religious Equality and Inclusive Development (CREID) which has been instrumental in supporting MRG on its journey to understand and counter hate speech. We also acknowledge the financial support from United Kingdom's Foreign, Commonwealth and Development Office (FCDO). The content is the sole responsibility of MRG. It does not represent the views of CREID and can under no circumstances be regarded as reflecting the position of the FCDO.

Table of Contents

About hate speech: what you need to know

Understanding, monitoring: what you can do and how

Responding, reporting: what you can do and how

Rebalancing content: what you can do and how

Pushing for reforms: what you can do and how

What you can't do (yet) and why not

Conclusion

Foreword

While monitoring the state of freedom of religion or belief around the world, in my capacity as the United Nations (UN) Special Rapporteur on freedom of religion or belief, I have witnessed first-hand the isolation, discrimination and even violence that peddling of hatred can engender. The fast-moving world of social media has served in particular to provide new spaces for old practices of intolerance to flourish, with digital content virality helping to spread polarizing speech to wider audiences than ever before. I have frequently reported on the alarming impacts, both online and offline, of incitement to hatred based on religion or belief, especially those that target persons facing multiple forms of vulnerability.

However, legislative restrictions of expression are often blunt and potentially even dangerous tools to address the dissemination of bigotry. States often use overly broad restrictions that infringe on speech that international law requires to be protected or that cause chilling effects on free speech. Laws often risk failing to strike the right balance between freedom of expression on the one hand and the right to non-discrimination on the other. Social media companies' attempts to apply content moderation to the phenomenon, meanwhile, have often been inscrutable, unevenly applied and of uncertain efficacy.

Restrictions have the impact of encouraging a 'balloon effect', whereby a clampdown on one set of platforms may simply cause a surge in advocacy of hatred in other, more permissive fora. More insidiously, speech constraints lead to an increase in 'borderline content', where malicious actors may contextually camouflage hatred in content specifically designed to evade moderation. While banning or prohibiting speech is occasionally necessary, particularly where it satisfies the Rabat Plan of Action's thorough six-part test to establish incitement, the shortcomings of prohibition demand alternative rights-based approaches, strategies, and tools

As emphasized by the landmark UN Human Rights Council Resolution 16/18, the most effective alternative response to 'hate speech' is counter-speech, while the 'Faith for Rights' framework stresses peer-to-peer learning as a useful means to address prejudice and stereotypes that drive hatred and fear. By addressing and rebutting problematic speech early, we can stop it from escalating into incitement to hostility, discrimination, or violence. I therefore heartily welcome the publication of this toolkit. Over the following pages, you will discover a detailed and thoughtful blueprint for human rights activists to engage with and tackle this phenomenon methodically. Whether you are concerned with understanding 'hate speech' better, gathering evidence about its extent, nature and consequences, working with platforms to reduce its spread, or distributing content to counterbalance the effects of advocacy of hatred, you will find helpful guidance here.

I hope that human rights defenders globally will make use of this toolkit to combat the escalating pattern of polarising and discriminatory speech, which threatens societal cohesion, mutual respect, justice and peace. While the immediate effects of advocacy of hatred are disproportionately borne by certain groups, conflict, disharmony and stalled development ultimately impact us all.

Dr Ahmed Shaheed

UN Special Rapporteur on freedom of religion or belief

Prefaces

If you stumbled upon this toolkit and did not seek it out, you may be wondering: 'Why hate speech? Why now?' After all, the climate crisis requires urgent global attention while human activities are increasingly pushed into precarity. But hate speech is one of the most virulent threats to good, effective and inclusive governance, which is essential for us to solve such complex issues.

Instead of climate change or decent work being central to decisions about who governs, power is being seized through scapegoat politics. Identity, wielded as a weapon, is used to generate artificial majorities (contrasted with 'others'), that can be whipped into a frenzy in a politics of distraction. As a formula it generated election wins in the United States, Myanmar, Poland, India, and others.

Messages of hate are now in the full glare of the public sphere. Individuals with no governance credentials have used these scare tactics to lever themselves into power, which they then use to hollow out long-standing democratic institutions and undermine the rule of law.

Hate speech is the weapon of choice. Ethnic, religious and linguistic minorities and indigenous peoples, as well as women, LGBTQ+ persons, people living with a disability, refugees and migrants, who can be singled out, bullied and blamed, have been the targets. The price paid by such communities has been catastrophic: harassment, intimidation, humiliation, destruction of properties, targeted physical violence, ethnic cleansing, crimes against humanity and even genocide. In the background, the chasm of structural discrimination, exclusion, and marginalization grows.

Faced with this, societies around the world are reconnecting with what is important. They are reaching out to 'the other' in the realization that thoes who seek to construct hate are often the sole (financial and political) beneficiaries.

This manual speaks to those people who thirst for change. To those with a determination to combat the hate among us. To those with the imagination and drive to work for a better world. To those motivated by the need to build open inclusive spaces for conversation, dialogue, mediation.

We trust you will find its analysis sharp and its suggestions useful. Most importantly, we trust that it will enable you to join with others who realize how education born from empathy remains the most effective tool in combating the spread of hate.

Wishing you luck and solidarity along the way,

Joshua Castellino

Executive Director
Minority Rights Group

This pioneering toolkit produced by Minority Rights Group and partners as part of the work of the Coalition for Religious Equality and Inclusive Development (CREID) is intended to be a resource for individuals, organizations and movements committed to stopping the mobilization of hate speech in ways that de-humanize, vilify and target people on account of their religious or non-religious beliefs.

Countering hate speech is at the core of CREID's mission. CREID was established with the view to 'redressing the impact of discrimination on the grounds of religion or belief, to tackle poverty and exclusion, and promote people's wellbeing and empowerment using research evidence and delivering practical programmes'.

There can be no progress made in improving people's wellbeing without tackling hate speech. Hate speech creates divisions in society that, in turn, create unequal and uneven opportunities for people to benefit from political, economic and social opportunities. Hate speech undermines the social cohesion of society, putting in jeopardy any improvements in human welfare through the negative impact of discrimination, insecurity and, in some cases, violence. As hate speech generates, amplifies and sustains a culture of 'us' vs 'them', humanity cannot flourish.

If the perpetrators of hate speech get away with inciting hate against an individual or group on the basis of their faith or not having a faith, or belief system today, who knows who they will target tomorrow? Anyone who is different from those who promote homogenizing ideologies may be next, no one can be fully assured that they will not be vulnerable to targeting.

This toolkit serves to highlight the dangers of hate speech against those of a faith or no faith, as well as providing practical ideas about how to challenge and counter the circulation of online hate. It is informed by the experiences on the ground of organizations working in very challenging settings, sharing lessons learnt from effective strategies but also reflecting on ongoing struggles challenging powerful actors and their agendas.

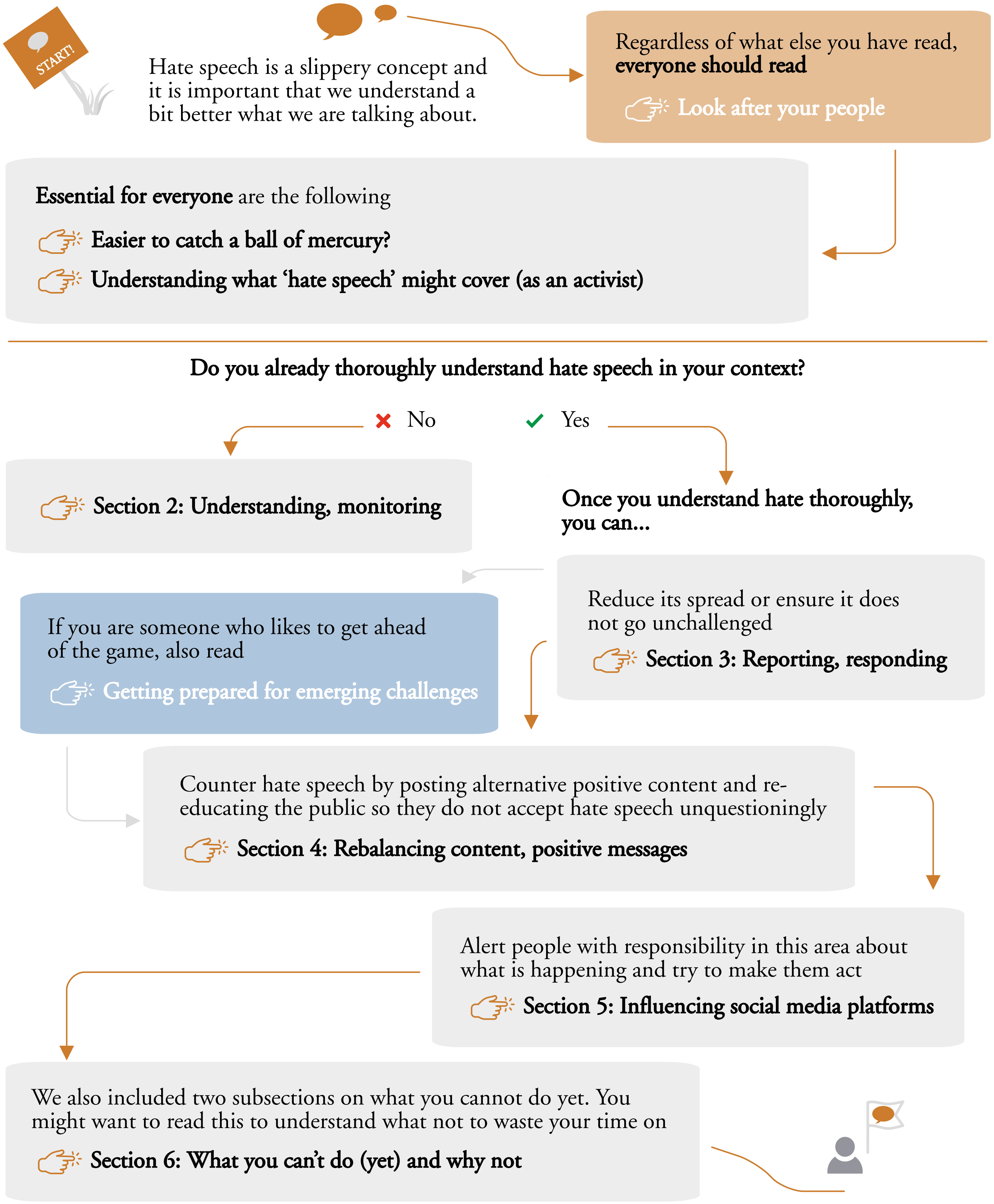
We hope that this toolkit will result in documentation and learning from your own experiences in countering hate speech, and this in turn will generate new iterative processes of learning and sharing.

Dr Mariz Tadros

Professor of Politics and Development
Institute of Development Studies, University of Sussex
Director, Coalition for Religious Equality and Inclusive Development

How to use this toolkit

We do not intend that you read this toolkit from start to finish. Instead you should feel free to dip in and out of its pages and read the sections that apply to you. A few sections are important for everyone but otherwise this annotated contents page aims to help you decide the most relevant content for your situation. We also included clickable external 🔗 and internal ☞ links.



Hate speech is a slippery concept and it is important that we understand a bit better what we are talking about.

Regardless of what else you have read, **everyone should read**

☞ **Look after your people**

Essential for everyone are the following

☞ **Easier to catch a ball of mercury?**

☞ **Understanding what 'hate speech' might cover (as an activist)**

Do you already thoroughly understand hate speech in your context?

✗ No ✓ Yes

☞ **Section 2: Understanding, monitoring**

Once you understand hate thoroughly, you can...

If you are someone who likes to get ahead of the game, also read

☞ **Getting prepared for emerging challenges**

Reduce its spread or ensure it does not go unchallenged

☞ **Section 3: Reporting, responding**

Counter hate speech by posting alternative positive content and re-educating the public so they do not accept hate speech unquestioningly

☞ **Section 4: Rebalancing content, positive messages**

Alert people with responsibility in this area about what is happening and try to make them act

☞ **Section 5: Influencing social media platforms**

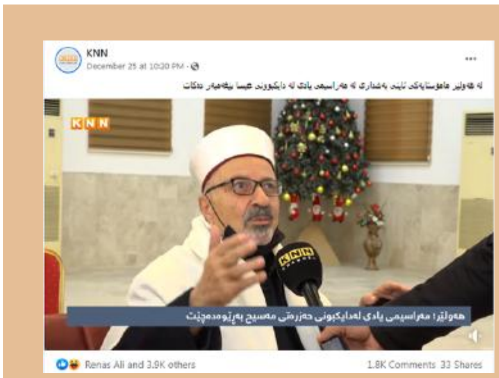
We also included two subsections on what you cannot do yet. You might want to read this to understand what not to waste your time on

☞ **Section 6: What you can't do (yet) and why not**

Hate speech in context

The world is experiencing an unprecedented rise in hateful speech. We see this offline, on our streets, even on bank notes, but also, increasingly, online on our screens. On the following pages we have compiled examples of experiences of hateful messaging and their ramifications, as well as comments about rising hateful expression. They show the impact on ordinary people of the daily onslaught of discriminatory hate-filled material, which in many cases directly incites violence.

Online


KNN
December 25 at 10:20 PM ·
لە هەولێر مامۆستایەکی ئاینی بەشداری لە مەراسیمی یادی لە دایکبوونی عیسا پێغەمبەر دەکات

Renas Ali and 3.9K others
1.8K Comments 33 Shares

Post engagement

👍 3.8k

💬 1.8k

33

Facebook post on a Kurdish page: *'In Erbil a Muslim imam visits a Christian church to celebrate the birth of Christ.'*

People were expressing torrents of hate against the moderate Muslim imam. Some accused him of being a kafir or an apostate because, in their view, he betrayed Islam by making this visit.

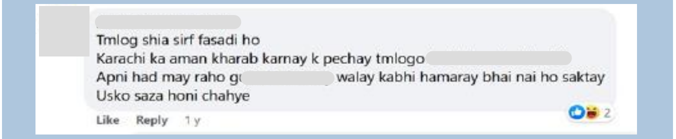
> *The people of Myanmar are easy to trust. They believe everything said by their trusted people, or inspired such as monks, leaders, influencers, celebrities, and community leaders. As a result, hate speech spreads quickly and is difficult to combat.'*

Monitoring volunteer, Peace Point Myanmar (PPM)

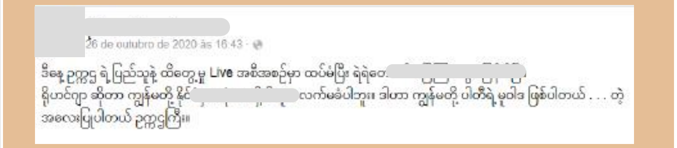
> *Social media penetration has given voices to the unheard, it has made them vulnerable too.'*

Haroon Baloch,
Bytes for All, Pakistan

In the Facebook conversation below, a user is not only demeaning the Shi'a community, but also abusing and inciting others to punish them.


[redacted]
Tmlog shia sirf fasadi ho
Karachi ka aman kharab karnay k pechay tmlogo [redacted]
Apni had may raho g[redacted] walay kabhi hamaray bhai nai ho saktay
Usko saza honi chahye
Like Reply 1 y
2

'*You Shi'a people are only troublemaking persons. You people are mainly involved in disrupting peace in Karachi. Be in your limits, one who commits blasphemy cannot be our brother, he must be punished.'*


[redacted]
26 de outubro de 2020 às 16:43 ·
ဒီနေ့ ဥက္ကဌ ရဲ့ ပြည်သူနဲ့ ထိတွေ့မှု Live အစီအစဉ်မှာ ထပ်မံပြီး ရဲရဲဝံ့[redacted]
ရိုဟင်ဂျာ ဆိုတာ ကျွန်မတို့ နိုင်[redacted] လက်မခံပါဘူး။ ဒါဟာ ကျွန်မတို့ ပါတီရဲ့ မူဝါဒ ဖြစ်ပါတယ် . . . တဲ့
အလေးပြုပါတယ် ဥက္ကဌကြီး။

Post by one of the People's Pioneer Party's candidates in 2020: *'Our Chairperson (Daw Thet Khine, PPP) answered the media question boldly. There is no Rohingya in our country. We don't accept the Rohingya. This is our Party's policy… Salute Chairperson.'*

In the media:

- *Al Jazeera* (2021)
- *The Diplomat* (2021)
- *The Friday Times* (2021)

'When the blood starts': Spike in Ahmadi persecution in Pakistan

Attacks and blasphemy cases against Pakistan's Ahmadi sect spike, driven by the rise of a far-right group and a religious campaigner.

THE DIPLOMAT

Pakistan's Social Media Is Overflowing With Hate Speech Against Ahmadis

The minority community has always faced persecution in the real world. Online, it's become even more common.

[redacted]
July 30, 2021

Social Media Brings Both Hope And Fear For Religious Minorities In Pakistan

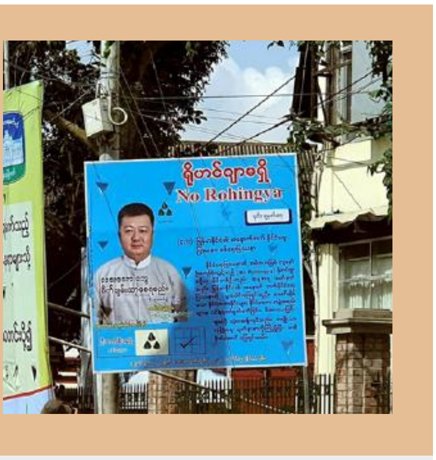
Digital media has become a double-edged sword for religious minorities by simultaneously being a source of unveiling majoritarian crimes, and a cesspool of hate speech against the vulnerable

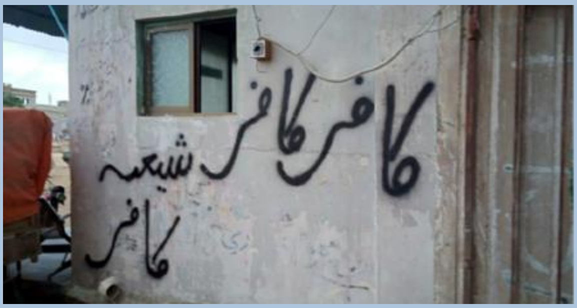
Offline

On 19 September 2021, an independent candidate, U Kyaw Soe Htut, who was running for a parliamentary seat in the Latha township in Yangon, Myanmar, used an anti-Rohingya slogan on his campaign posters saying ‘NO Rohingya’.

On the poster in the picture on the right, it says:

‘I, Kyaw Soe Tun, determined there are no Rohingya. I will cooperate in solving the political issues of the west gate of Myanmar [referring to Rakhine state and Bangladesh border and Rohingya issues]. *I oppose interference by unrelated countries and international organizations. It will be addressed in the face of national security.’*





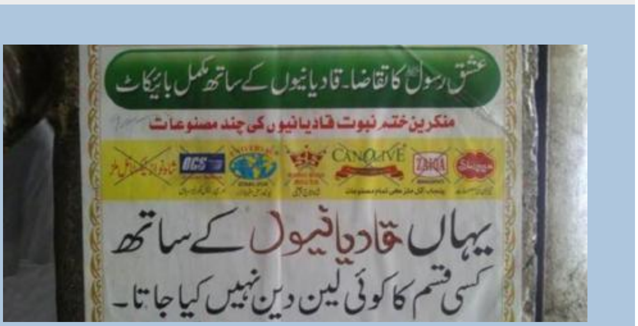
Graffiti on a wall against the Shi’a community saying *‘Shi’as are infidels.’*

> **Even though the MPs spoke hatefully during their election campaign, the union election commission did not prohibit them, didn’t take action. And so, hate speech spreads easily and leads to violence.’**

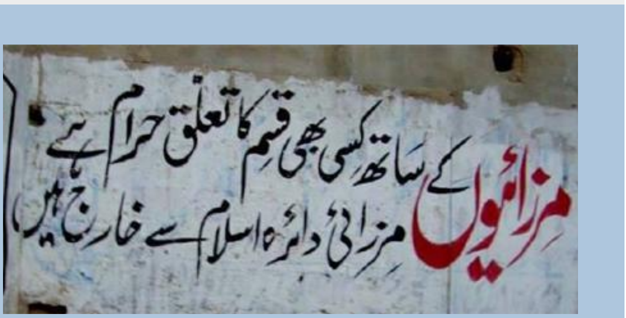
Monitoring volunteer, PPM, Myanmar

> **Although, Christian community has served and is serving Pakistan in the fields of education and health, we still experience hate-speech in almost every profession.’**

Christian nurse from Lahore, Pakistan


عشق رسول کا تقاضا۔ قادیانیوں کے ساتھ مکمل بائیکاٹ
منکرین ختم نبوت قادیانیوں کی چند مصنوعات
یہاں قادیانیوں کے ساتھ
کسی قسم کا کوئی لین دین نہیں کیا جاتا۔

A poster outside a shop in Rawalpindi, Pakistan, saying, *‘No economic transaction is allowed with Ahmadiyya people.’*


مرزائیوں کے ساتھ کسی بھی قسم کا تعلق حرام ہے
مرزائی دائرہ اسلام سے خارج ہیں

A wall chalking in Karachi, 2021, saying *‘Any relation with Ahmadiyya people is forbidden and they are not part of Islamic community.’*

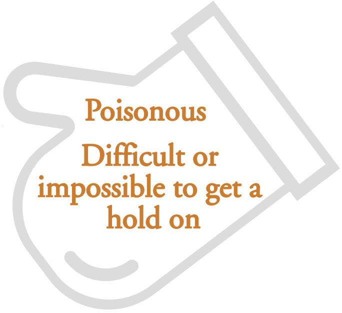
> **People do not know much about the severity of hate speech. In our mind, hate speech is using bad words for people but this is not the case. Today, I got to know about what actually is hate speech.’**

Female attendee at training workshop, Pakistan

About hate speech
what you need to know

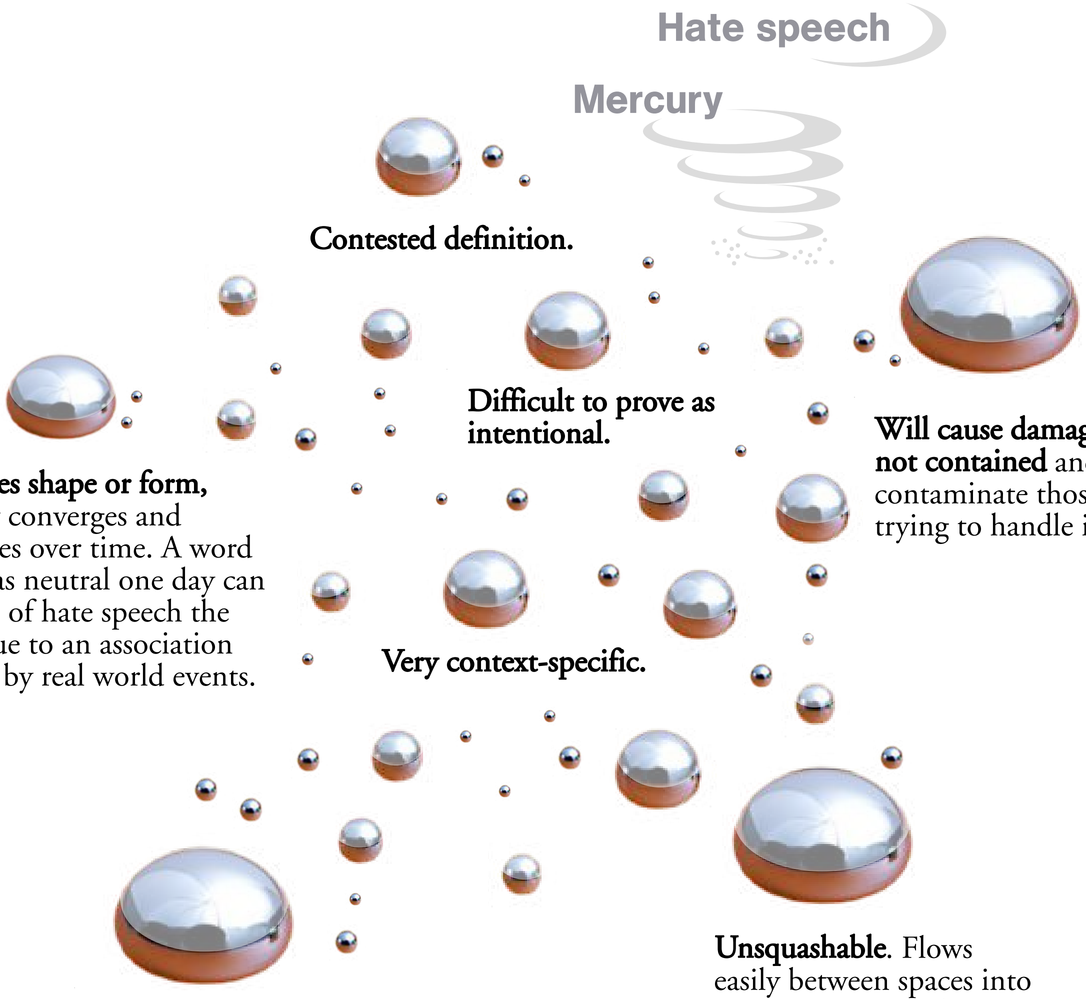
1.1 Easier to catch a ball of mercury?

> *Hate speech has been so prevalent in the society that now it has become normal speech.'*
>
> Participant of the focus group discussion
> Hindu community member, rural Sindh



Poisonous

Difficult or impossible to get a hold on



Contested definition.

Difficult to prove as intentional.

Will cause damage if not contained and may contaminate those trying to handle it.

Changes shape or form, rapidly converges and disperses over time. A word that was neutral one day can be part of hate speech the next due to an association caused by real world events.

Very context-specific.

Best approached with guidance and side-wise, not through a direct attack. To pick up a ball of mercury you need tools – a pipette, a testtube and a cork – and to create a force or a vacuum that acts against it.

Unsquashable. Flows easily between spaces into cracks. Hate speech moves across online platforms instantaneously. It also moves between the digital and real world.

1.2 Understanding what 'hate speech' might cover (as an activist)

As you can see from examples of what may be seen as hate speech (pp. 7–8), there are many types of expression that might be considered 'offensive speech' or 'hate speech' – from an insult to the use of negative stereotypes in a campaign poster, from a call to violence against a group to an article justifying discrimination. So, pinning down one singular definition that everyone agrees covers all forms of 'hate speech' is a monumental task. It is important to state at the outset that not all hate speech should be prohibited or restricted, and here we are talking about hate speech in general.

Why is it so difficult to define 'hate speech'?

As the table below shows, the exact same set of words can be considered damaging hate speech or not depending on the context, the speaker, the speaker's intention and the potential impact. This is part of the reason why settling on a final universally agreed definition has proved difficult.

Importance of context

Whether or not an expression is intended and/or perceived as hateful heavily depends on the context it takes place in.

Difficulty in real life: Context can be subjective – was the expression really a joke, as a speaker might claim, or not?

Importance of the potential or actual impact

To be considered 'hate speech', the expression must have a potential real-world impact, either causing actual harm or making it likely that actual harm will occur. In our example, saying 'We should just kill all followers of the Alpha-Centauri religion' while alone in a closed room would not be considered hate speech.

Difficulty in real life: Did an impact take place at all? How much time needs to pass to be sure? What if steps were taken to mitigate an impact, who is to say if it would have happened or not? If the harm is limited to discrimination, how do we prove it? And even if impact is apparent, how direct is the link between it and the hateful expression really?

Importance of intention

Many people consider that speech must be intended to cause harm to the target. If someone says something entirely innocent which is widely interpreted to convey hate, and it is clear that this is an honest mistake or misunderstanding, that is not hate speech even if it directly leads to harm.

Difficulty in real life: How do you prove that someone intended their damaging speech to cause harm? What should the threshold for such proof be?

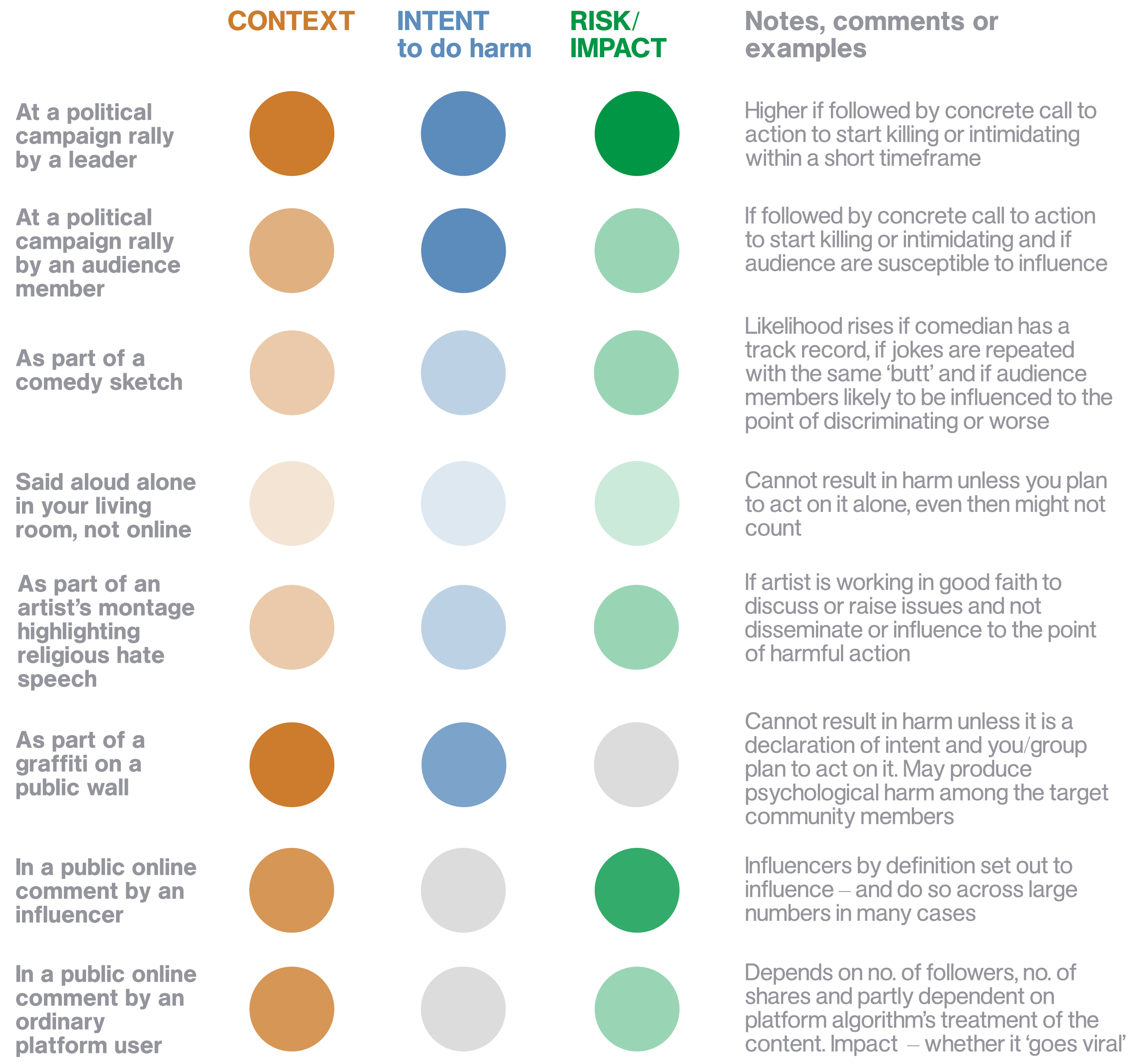
The danger of misusing the term 'hate speech'

Be careful not to use the term 'hate speech' too widely. This might water down the seriousness of hate speech as a phenomenon in the perception of your audience, which would make your work less effective. Down the line, the term 'hate speech' can then lose some of its original meaning. That can help your government use the fight against supposed 'hate speech' as a pretext for passing legislation restricting legitimate forms of expression.

? Is ‘We should just kill or banish all followers of the Alpha-Centauri religion’ ‘prohibited or restrictable hate speech’ depending on the context/author and intent?

Very likely ● ● ● ● ● ● ● ● ● ● Very unlikely

● Difficult to assess but consider track record and knowledge of context



	CONTEXT	INTENT to do harm	RISK/ IMPACT	Notes, comments or examples
At a political campaign rally by a leader	●	●	●	Higher if followed by concrete call to action to start killing or intimidating within a short timeframe
At a political campaign rally by an audience member	●	●	●	If followed by concrete call to action to start killing or intimidating and if audience are susceptible to influence
As part of a comedy sketch	●	●	●	Likelihood rises if comedian has a track record, if jokes are repeated with the same ‘butt’ and if audience members likely to be influenced to the point of discriminating or worse
Said aloud alone in your living room, not online	●	●	●	Cannot result in harm unless you plan to act on it alone, even then might not count
As part of an artist’s montage highlighting religious hate speech	●	●	●	If artist is working in good faith to discuss or raise issues and not disseminate or influence to the point of harmful action
As part of a graffiti on a public wall	●	●	●	Cannot result in harm unless it is a declaration of intent and you/group plan to act on it. May produce psychological harm among the target community members
In a public online comment by an influencer	●	●	●	Influencers by definition set out to influence – and do so across large numbers in many cases
In a public online comment by an ordinary platform user	●	●	●	Depends on no. of followers, no. of shares and partly dependent on platform algorithm’s treatment of the content. Impact – whether it ‘goes viral’

An intention to incite is crucial when determining the culpability of someone using a hateful expression. Keep in mind when going through the rest of this section that legally regulating those who use hate speech will depend on establishing their harmful intention. Not doing this would be an abuse of the right to freedom of expression. This is one of the reasons why many human rights organizations do not advocate for prohibiting most, let alone all, forms or occurrences of ‘hate speech’. Taking a context-specific approach, providing policy recommendations, countering the underlying causes of hate speech in your society and dealing with its impacts may be more important than the legal definition or debating the intentions of the speaker.

Hate speech and the law

The issue of hate speech raises complex legal questions that vary depending on the laws applicable in each country as well as the context in each country.

'Hate speech' is broadly defined by the United Nations (UN) 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.'*

Non-legally binding working definition

Note that incitement to **discrimination** is included along with incitement to hostility or violence.

The general provision of Article 20, of the International Covenant on Civil and Political Rights (ICCPR) 1966 obliges states to prohibit: *'[A]ny advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence.'*

Article 19 of the ICCPR covers freedom of speech but notes in sub-section 3 that *'The exercise of the rights ... may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary:*
(a) For respect of the rights or reputations of others;
(b) For the protection of national security or of public order (ordre public), or of public health or morals.'

See what your state has signed up to

To clarify state obligations with regard to hate speech, the UN further developed the Rabat Plan of Action (2012) which puts forward a six-part test for hate speech. If these conditions are met in any given situation, the Rabat Plan of Action calls on states to institute a system of criminal sanctions.

Unlike a treaty that is binding on states that are party to it, the Rabat Plan of Action is non-binding, but it can still be useful in helping us unpack the concept of hate speech.

Any type of hateful expression that is **not specifically prohibited** under an international treaty or convention or **lawfully restricted** by a given state is likely to fall under the right to freedom of expression under international law. This of course does not mean such expressions are **protected from criticism** – but you should keep this in mind when campaigning against hate speech in your context.

Six-part test for hate speech

Is discrimination, hostility, or violence resulting from the hateful expression likely because of:

□ the social and political context?

□ the ability of the speaker to influence their audience?

□ an intent to promote hatred publicly (as opposed to just being reckless)?

□ the content (what was said) and the form (how it was said) – were they provocative and direct?

□ the extent of the audience: was the expression addressed to a large audience, or an audience prone to follow incitement?

□ a reasonable probability that the expression will result in inciting harm for or action against the group?

> *Hate speech creates an us-against-them attitude which is very difficult to get rid of.'*
>
> Baha'i survey participant, Iraq

Legal provisions addressing hate speech at a glance

There are two other key UN treaties that you should be aware of:

INTERNATIONAL CONVENTION FOR THE ELIMINATION OF ALL FORMS OF RACIAL DISCRIMINATION 1965

Article 4, requires States to: *'condemn all propaganda and all organizations … which attempt to justify or promote racial hatred and discrimination in any form'*. States are required to take immediate and positive measures: *'to eradicate all incitement to, or acts of, such discrimination'*. This includes declaring: *'all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination'* as offences punishable by law. The text makes clear that this includes any: *'group of persons of another colour or ethnic origin,'* and ethnicity commonly intersects with religion. 🔗

CONVENTION ON THE PREVENTION AND PUNISHMENT OF THE CRIME OF GENOCIDE 1948

Punishable acts under Article III include: *'Direct and public incitement to commit genocide;'*. 🔗

Many other institutions have delineated approaches towards dealing with hate speech over the years, many in regional systems. It is good to be aware of these and some may be useful in building pressure on your government, if your state is part of these bodies.

AFRICAN UNION

The African Charter on Human Rights and Peoples' Rights in Article 9(2): restrictions on rights are permissible as long as they are *'within the [domestic] law'*. 🔗

BUT: Declaration of Principles on Freedom of Expression in Africa, Article 13(2) – freedom of expression should not be restricted *'unless there is a real risk of harm to a legitimate interest and there is a close causal link between the risk of harm and the expression'*. 🔗

COUNCIL OF EUROPE

Committee of Ministers Recommendation CM/Rec (1997) 20 prohibits: *'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 against minorities, migrants and people of immigrant origin'*. 🔗

INTER-AMERICAN CONVENTION ON HUMAN RIGHTS

Article 13, para. 5 states: *'Any propaganda for war and any advocacy of national, racial, or religious hatred that constitute incitements to lawless violence or to any other similar illegal action against any person or group of persons on any grounds including those of race, color, religion, language, or national origin shall be considered as offenses punishable by law.'* 🔗

ORGANISATION OF ISLAMIC COOPERATION

The Cairo Declaration of the Organisation of Islamic Cooperation on Human Rights in Article 21 on the Right to Freedom of Opinion and Expression states: *'b. Everyone shall have the right to freedom of expression.'* with restrictions on the exercise of this right limited to *'Advocacy of hatred, discrimination or violence on grounds of religion, belief, national origin, race, ethnicity, color, language, sex or socio-economic status.'* For full text see: 🔗

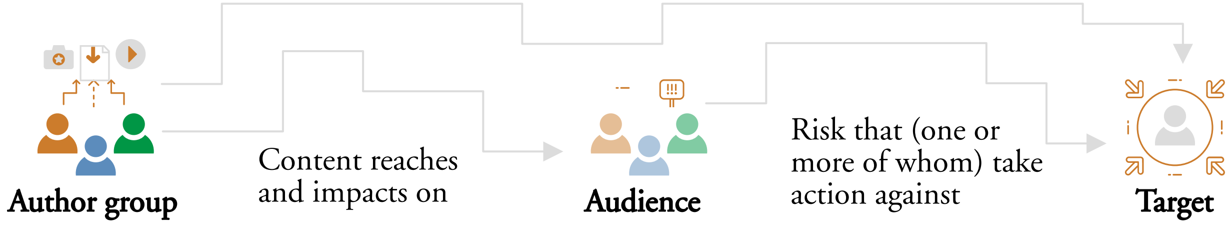
You will need to learn about the details of national hate speech and free speech regulations in your country. But remember: many types of hateful or distasteful expression that you may encounter daily are protected speech under international law. The right to freedom of expression is a fundamental human right, and legal restrictions on it (prohibition, criminalization, censorship) are easily abused to the detriment of human rights, and especially minority rights.

What is NOT generally considered prohibited hate speech?

- Rudeness, insults, cricitism, disrespect, swearing, disagreement, intolerance, put downs, political opinions, expression deemed blasphemous – none of these on their own are enough to constitute prohibited hate speech (even when they are directed towards those of one faith or no faith).
- Content which is negligent or careless or reckless is not prohibited hate speech if the person who produced it had no intention of causing or contributing to discrimination/hostility/violence (even if they did so, as long as it was purely by accident).
- Blatantly hate-filled and hate-motivated speech is not prohibited hate speech if there is no reasonable prospect that it could lead to, result in or somehow directly contribute to discrimination, hostility or violence against a particular group.
- Criticism, satire or ridiculing of religions or beliefs, rituals and traditions, statements like ‘my religion is better than yours’ or ‘people who follow religion A are wrong or misguided because …’, unless it incentivizes discrimination, hostility, or violence against all followers of that religion.

Why are minorities particularly vulnerable to hate speech?

Hate speech can be seen as a process involving three actors:


Author group
Content reaches and impacts on
Audience
Risk that (one or more of whom) take action against
Target

The personal risk resulting from hate speech to any given member of society is fairly equal up until the risk of action. Compared to majorities in societies, minorities who are marginalized, often poor and politically weak are easy targets, less able to defend themselves, and it is less likely that the state (or anyone) will intervene on their behalf.

> *Hate speech simply makes victims feel insecure about their identities and makes them feel like low-class people; they feel discriminated against, suppressed, shy, uncomfortable, and vulnerable about themselves. It may damage their mental health, and destroy their living lives.’*
>
> Htet Swe, PPM, Myanmar

MRG therefore strongly believes that it follows logically that hateful expression against minorities should be **much** more likely fall into the area of speech that must be prohibited when everything else is equal. The risks of people acting on hate-filled content are higher where minorities are concerned because of the power imbalance between them and the majority group where they live. Read more about the current state of minority persecution on social media in the 2021 report of the UN Special Rapporteur on minority issues.

1.3 Digital–real-world exchange and flow

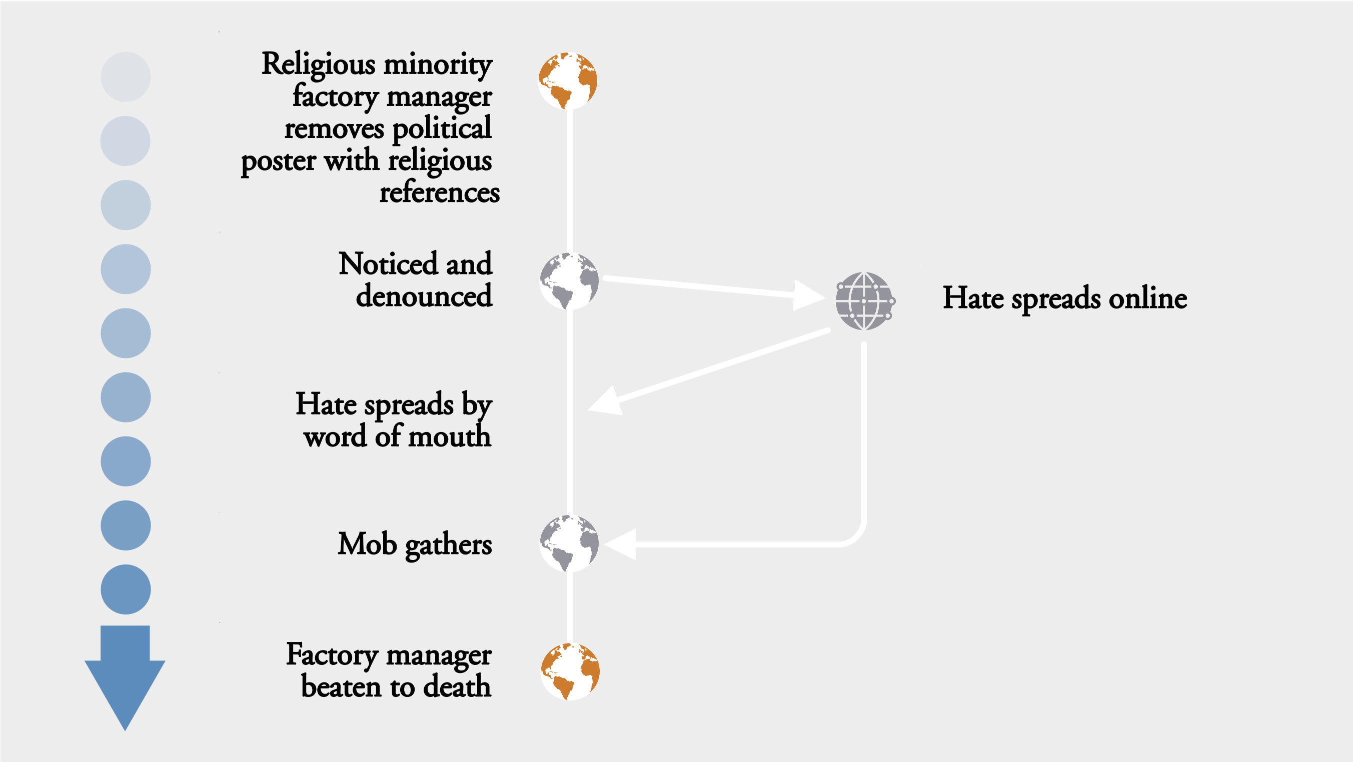
Public debate and personal interactions in many societies worldwide are increasingly taking place both in the physical world and in the digital realm. Events happening in the real world will often end up on the internet. Meanwhile, the content we consume digitally influences our real-world lives to a certain extent. This constant exchange and flow of information between the two realms also has a direct role in amplifying diffferent kinds of hateful expression. A single person’s act of hate now has the potential to reach and incentivize thousands, if the algorithm enables this: ☞ See section

> ‘*Spewing hate in online spaces can have a manifold impact. At times it becomes a dangerous tool in the hands of the majority against minority groups. We have witnessed events of mob lynching or violence against religious groups in offline spaces that originally started online.’*
>
> Activist, Islamabad, Pakistan

> ‘*In 2020, I think there are groups who are behind the scenes creating messages of hate speech against religious minorities online and offline. These people are encouraging brainwashing and politicians are spreading these messages online and offline.’*
>
> Monitoring volunteer, PPM, Myanmar

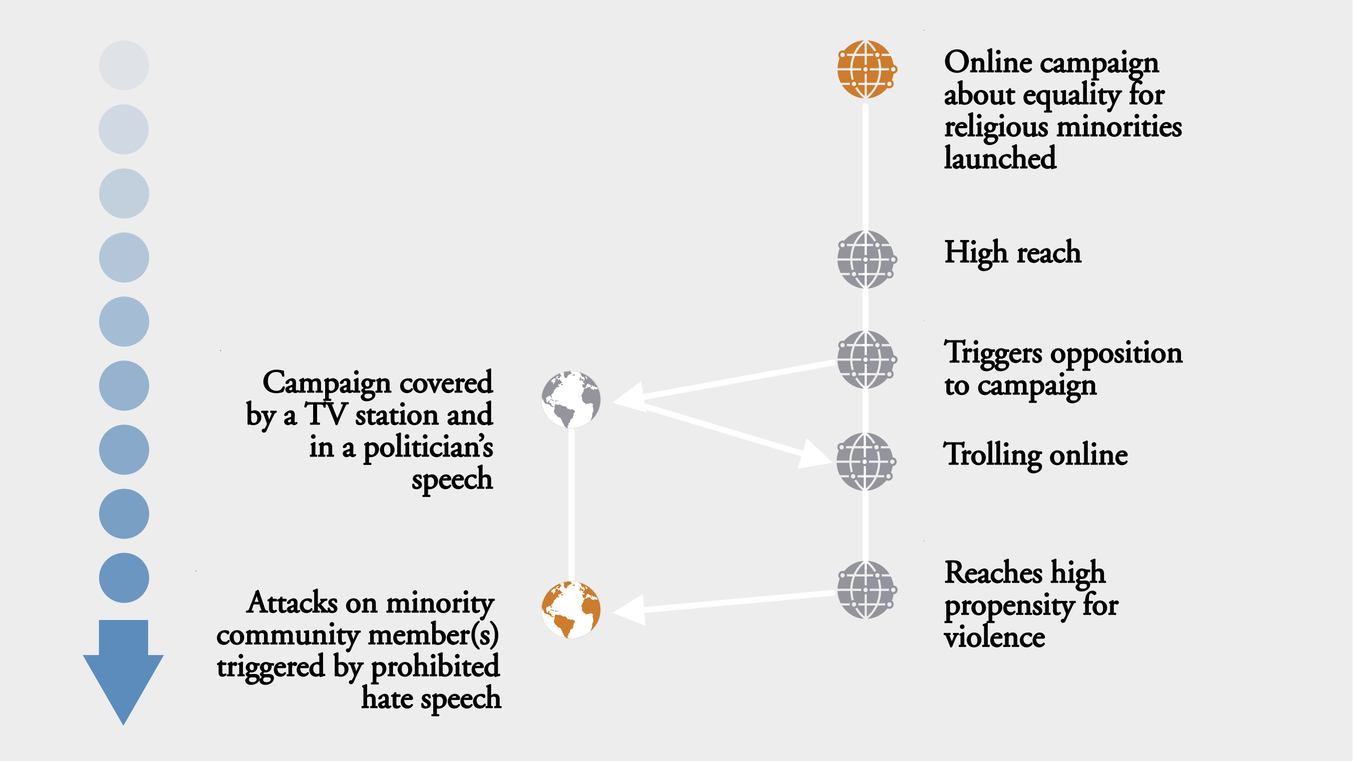
Below are three general examples of how prohibited hate speech flows between the two dimensions.

1 Beginning in the real world



Religious minority factory manager removes political poster with religious references
Noticed and denounced
Hate spreads online
Hate spreads by word of mouth
Mob gathers
Factory manager beaten to death

2 Beginning in the digital realm



Online campaign about equality for religious minorities launched

High reach

Triggers opposition to campaign

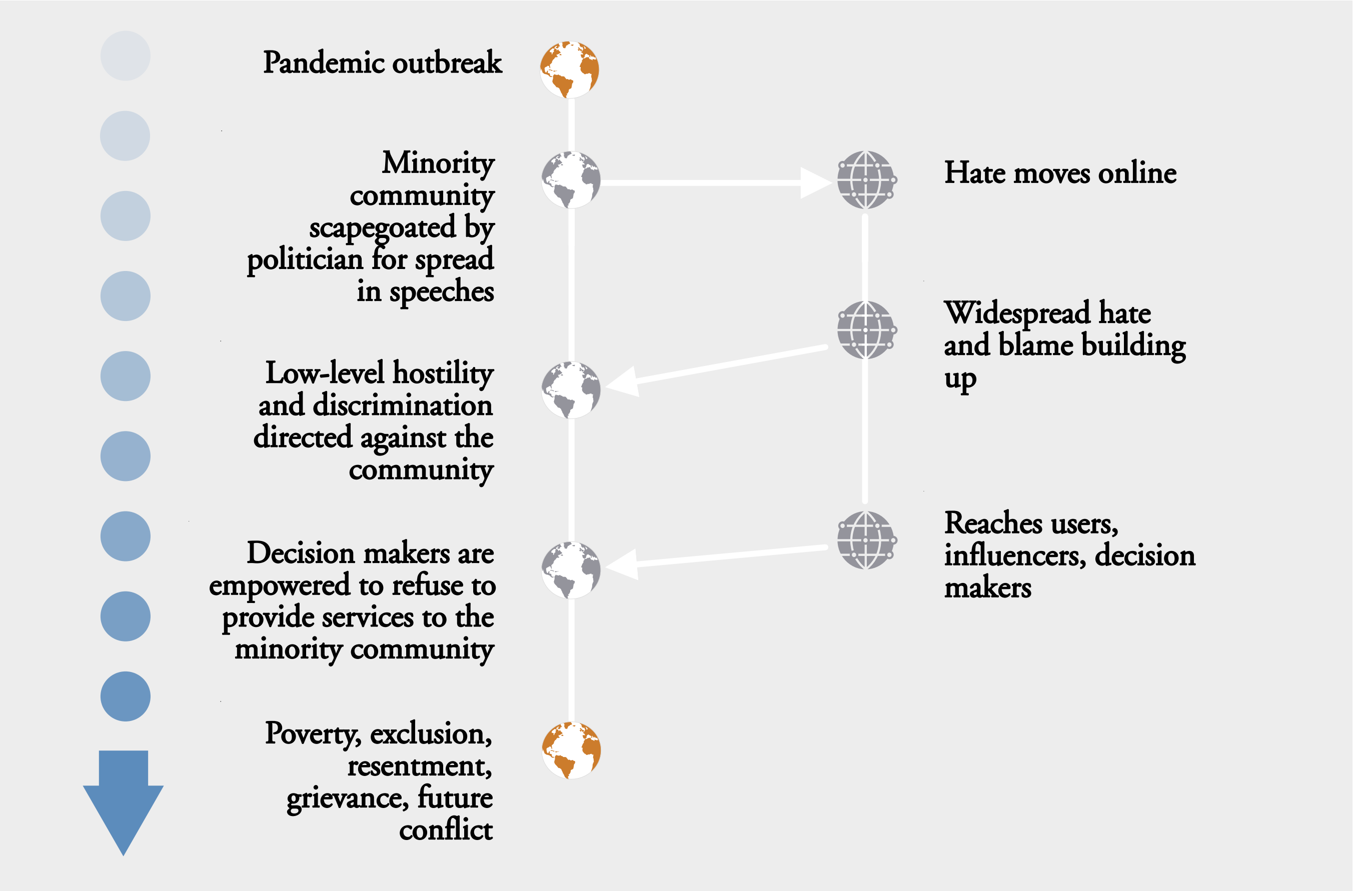
Campaign covered by a TV station and in a politician's speech

Trolling online

Reaches high propensity for violence

Attacks on minority community member(s) triggered by prohibited hate speech

3 Hate speech resulting in discrimination



Pandemic outbreak

Minority community scapegoated by politician for spread in speeches

Hate moves online

Widespread hate and blame building up

Low-level hostility and discrimination directed against the community

Reaches users, influencers, decision makers

Decision makers are empowered to refuse to provide services to the minority community

Poverty, exclusion, resentment, grievance, future conflict

Understanding, monitoring

what you can do and how

2.1 Creating a lexicon of hateful terms

Case study

How the Catholic Commission for Justice and Peace researched and published a lexicon of hateful words

Sharoon Sharif

In 2018, CCJP decided to work on a lexicon of hateful terms used against or experienced by religious minorities because we were witnessing widespread and increasing hate in Pakistan, which marginalized religious minorities and was contributing to an increasingly polarized society.

> *‘Why would we want to compile a full list of all hateful terms? Because many of them are used daily and not everyone using them or hearing them may know that they are offensive.’*

We started by mapping out the religious communities experiencing hate in our country, reading about their situation, problems and experiences. Although we ourselves are already a religious minority, we wanted to be well informed about and respectful to those of other faiths or no faith at all times.

We organized many focus group discussions (FGDs) with different religious groups including Ahmadis, Christians, Hindus, Shi’a and Sikhs (144 in Sindh, 71 in Khyber Pakhtunkhwa, and 126 in Punjab, in total 341 people). We met with people in groups according to their religion so that there was a safe space to discuss this sensitive topic. Our plans included holding an FGD with atheists. This group is very sensitive in Pakistan but it was difficult to build up trust between them and us. Our Christian roots also did not help. We tried, via trusted intermediaries, to organize a group but ultimately relied on input from some individuals (some outside Pakistan).

All the words mentioned to us that people considered demeaning or offensive were compiled into one database; we used Microsoft Excel. We researched each word in detail, by speaking to religious minorities and tracking the etymology of each word. We decided to group the terms which affected each community together, so that each religious group had a section in the lexicon.

Catholic (National) Commission for Justice and Peace (CCJP)

Human rights body, established by the Pakistan Catholic Bishops’ Conference in 1985.

Advocacy organization focusing on the human rights of the marginalized, especially religious minorities, women, children and labour in Pakistan, which involves interventions regarding awareness and opinion building about law and policy reforms.

NCJP NATIONAL COMMISSION FOR JUSTICE AND PEACE

Find out more

Within each section, we decided to order the terms alphabetically, as they are pronounced. This means it is easy to look up a particular term and find out about it. Our lexicon included Urdu and Roman scripts to be useful to different audiences. The draft was shared with the focal people of each community to proofread and check.

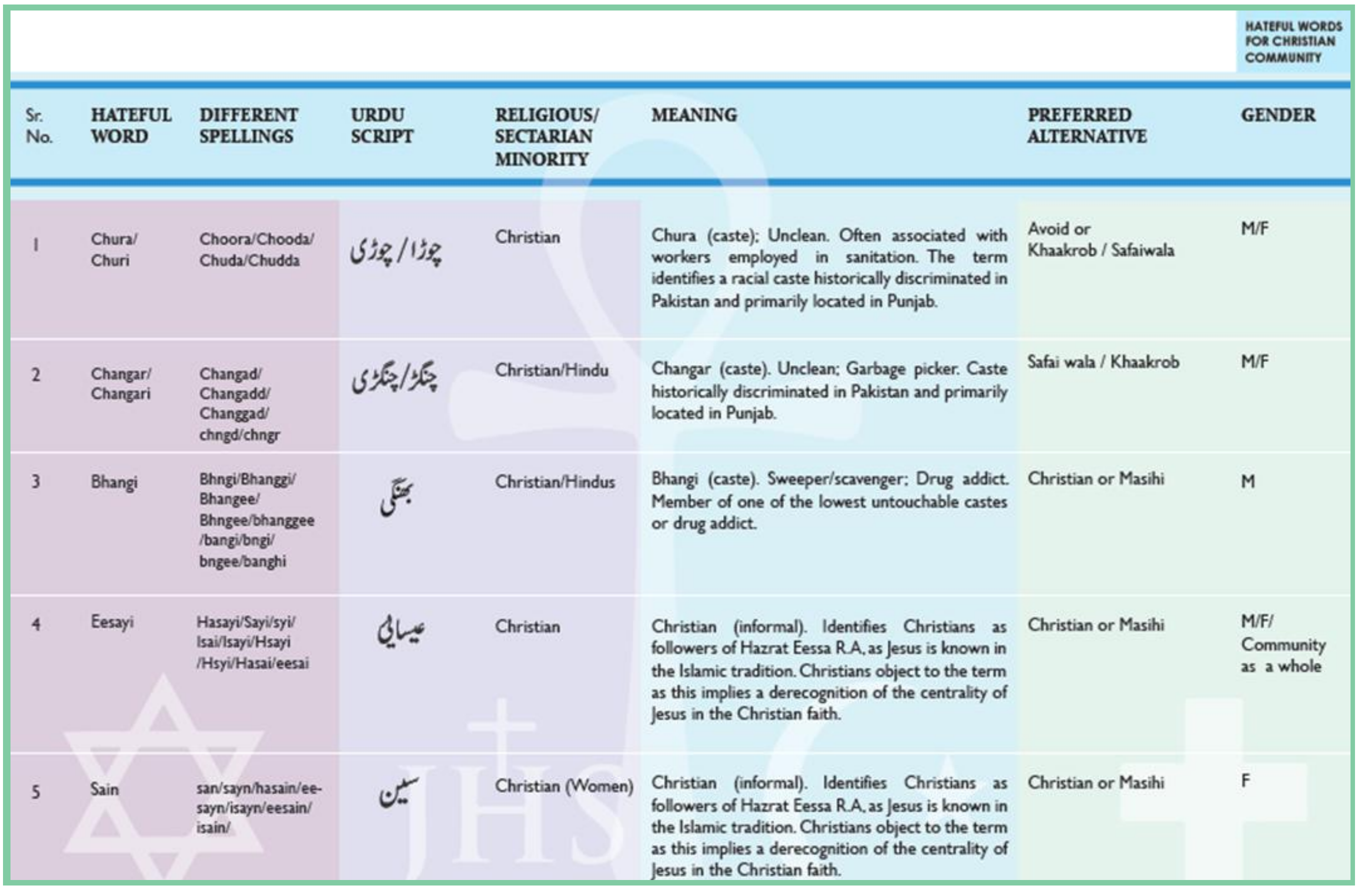
After designing, proof-reading and printing, we shared the lexicon with a small number of like-minded people (e.g. social media platform country teams, human rights NGOs, diplomatic missions and international agencies). We knew that there were risks to making the Lexicon generally available to the public due to religious intolerance and hostility. The Lexicon was useful to our partners Bytes for All (B4A) for creating monitoring queries and Bargad for training youth activists. The process of researching the lexicon also informed a series of policy briefs that CCJP produced, seeking to influence provincial-level politicians to take action, reduce or counter all forms of hate speech. A female MP from Punjab said that the lexicon is a very practical document to reduce hate speech if it is used in a positive manner, particularly by print, electronic and social media outlets.

We learnt that in general our entire society was ignorant of the prevalence, internalization, and severity of consequences of the many forms of hate against religious minorities, including the minority communities themselves. Because many religious minorities experience hate every single day, they get very used to it, and in fact, although it continues to take a toll, they can almost stop noticing it at a conscious level.

For reasons explained above, we did not share the link to an online copy of the lexicon here. If you would like to receive a copy of the Lexicon, you can write to CCJP via their website and explain clearly who you are and what you want to do with it.

Contact CCJP

Screenshot of the lexicon of hateful terms. Courtesy of the Catholic Commission for Justice and Peace



HATEFUL WORDS FOR CHRISTIAN COMMUNITY

Sr. No.	HATEFUL WORD	DIFFERENT SPELLINGS	URDU SCRIPT	RELIGIOUS/ SECTARIAN MINORITY	MEANING	PREFERRED ALTERNATIVE	GENDER
1	Chura/ Churi	Choora/Chooda/ Chuda/Chudda	چوڑا / چوڑی	Christian	Chura (caste); Unclean. Often associated with workers employed in sanitation. The term identifies a racial caste historically discriminated in Pakistan and primarily located in Punjab.	Avoid or Khaakrob / Safaiwala	M/F
2	Changar/ Changari	Changad/ Changadd/ Changgad/ chngd/chngr	چنگڑ/چنگڑی	Christian/Hindu	Changar (caste). Unclean; Garbage picker. Caste historically discriminated in Pakistan and primarily located in Punjab.	Safai wala / Khaakrob	M/F
3	Bhangi	Bhngi/Bhanggi/ Bhangee/ Bhngee/bhanggee /bangi/bngi/ bngee/banghi	بھنگی	Christian/Hindus	Bhangi (caste). Sweeper/scavenger; Drug addict. Member of one of the lowest untouchable castes or drug addict.	Christian or Masihi	M
4	Eesayi	Hasayi/Sayi/syi/ Isai/Isayi/Hsayi /Hsyi/Hasai/eesai	عیسائی	Christian	Christian (informal). Identifies Christians as followers of Hazrat Eessa R.A. as Jesus is known in the Islamic tradition. Christians object to the term as this implies a derecognition of the centrality of Jesus in the Christian faith.	Christian or Masihi	M/F/ Community as a whole
5	Sain	san/sayn/hasain/ee-sayn/isayn/eesain/ isain/	سین	Christian (Women)	Christian (informal). Identifies Christians as followers of Hazrat Eessa R.A. as Jesus is known in the Islamic tradition. Christians object to the term as this implies a derecognition of the centrality of Jesus in the Christian faith.	Christian or Masihi	F

10 steps to...

Creating a lexicon of hateful terms

1 **Check** what already exists, consult experts, conduct a thorough literature review. Forewarn and consult all minority communities you know and make sure they agree

2 **Make sure** that you have a thorough understanding of differences between minority and majority communities and local debates. Define who is covered and make sure everyone within that definition is included

3 **Consider** a full range of words, e.g. words not just for people but also actions, objects or places that are associated with the minority groups

4 **Conduct** focus group discussions and interviews with at least 30-50 members from each minority covered in your lexicon. Explain different definitions of hate speech, ask for examples and avoid mixed group meetings

5 **Ensure** that you involve men and women (and others) of different social classes, ethnicities, ages and political affiliations, and people with disabilities as far as possible, and a wide range of locations and occupations

6 **Document** words groups say are offensive, research each term, understand why it is hurtful for the community and what are their preferred terms to use instead

7 **Ask** older people within the minority community where certain terms have come from, if you can't find it online or in studies. Sometimes it helps if people understand the origin of a hateful term

8 **Show** the compiled data to each community's focal people for proof-reading and also check again their preferred terms in place of hurtful wordings

9 **Consider** responsibly sharing it with allies and neutral people who will use it without fuelling hate

10 **Use** to monitor all kinds of hateful content online or make it available to like-minded media outlets and social media platforms to help in their work

2.2 Gathering data manually

Case study

How Independent Media Organisation in Kurdistan gathered hate data without using custom made software

Pshtiwan Faraj Mohammed

We organized consultation sessions with key stakeholders from all religious groups in Iraq to make them aware of our plan, and to ensure synergies with existing initiatives. Four consultation sessions took place in Erbil, Mosul and Baghdad, and two additional meetings with the Sunni Endowment and the Shi'a Endowment, respectively. The participants were asked to identify the forms of hateful expressions and the terms and phrases used to denigrate their communities they know about. Gender inclusion was mainstreamed throughout.

We formed a specialized team of three to monitor adverse content affecting religious communities on social media platforms, who, between them, spoke Kurdish, Arabic and English. We trained the team on the different kinds of hate speech, vocabularies and patterns to look for, and how to measure the impact and spread of hateful content. The online content they identified was systematically collected in a database to allow for further analysis and documentation.

Building on the results of social media monitoring, the project team released periodic bulletins with the main trends and findings. The bulletins were drafted in-house in cooperation with the Iraq Media House (an Iraqi NGO specialized in media monitoring). Gender being important, speech targeting women was included in the research and the bulletins.

A series of three trainings for 45 journalists on responsible reporting were organized in Basra, Baghdad and Erbil. The trainings covered many forms of religious hate speech, its consequences, and how to avoid common pitfalls when reporting sensitive content and stories about ethnic and religious differences. The training materials were based on the social media monitoring activity and included good and bad practices, as well as their real-life consequences. Trainees were encouraged to report content in a way that exhibits critical thinking, fosters understanding, and resists sectarian narratives.

Independent Media Organisation in Kurdistan

NGO in northern Iraq founded in 2008 by Free Press Unlimited, working to improve the quality of Iraq's media and to promote peace, tolerance and coexistence and advocate for marginalized communities.

Target audience are women, youth, IDPs, refugees, religious minorities in Iraq, journalists and CSO activists.

IMOK INDEPENDENT MEDIA ORGANISATION IN KURDISTAN

Find out more

After a two-day workshop on grassroots advocacy campaigns, the project produced a series of advocacy materials: 612 brochures (presenting statistics and findings from the project, in Arabic and Kurdish) and 260 handbooks (serving as a guide to practitioners, in Arabic and Kurdish). In addition, the campaign materials also called for the enactment of laws against internationally prohibited hate speech and were distributed to policymakers in Baghdad and across Iraq.

10 steps to...

Gathering data manually (to inform policy, advocacy and reform)

1 **Consult** and seek engagement with key stakeholders, community and religious leaders, on the process and get their support

2 **Consult** with all the potential targets of hate in your area about the terms that they find offensive and the discrimination, hostility and violence that they experience. If you do not have access to a lexicon yet:
☞ See section

3 **Design** a database to collect all the relevant characteristics linked to hateful expressions, e.g. target (religion, gender, age, livelihood, other characteristic if any), author type (e.g. politician, religious leader, celebrity), author gender, publication media, reach

4 **Record**, if possible, whether it was reported and taken down and, if so, after how long. Decide if you are interested in time of day, geography or other relevant factors that you might be able to identify when scanning content; every element captured adds to the time taken

5 **Create** and train your team on definitions of hate speech relevant in your context, vocabularies, patterns to look for and your database

6 **Deploy** your team to read online content, identify hate speech content based on your working definition, log and categorize the relevant content on your system

7 **Ask** the team, if time allows, to follow up on the identified hate-filled content breaching the content policy of the relevant platform to see if it is taken down and if so, after how long

8 **Use** your database to identify and analyse the main targets, authors and media spreading hate and the numbers of people viewing or sharing it

9 **Share** your results with social media companies, human rights monitoring bodies in your country and internationally

10 **Look after** your people
☞ See section

2.3 Gathering data using software

What you need to know

Semi-automated data collection using social media monitoring software tools

Haroon Baloch

Gathering data manually (for policy, advocacy and reform) can be slow, labour intensive and therefore costly. Here we show you the general design and some functions of monitoring software tools which allow you to both collect and understand relevant data. You will need to check a sample of the results to be sure that what is captured is accurate. The software tools won't do all the work for you, you will need to work on the extracted data to understand what is going on. What the software tools do is get you a smaller set of mostly relevant data (with linked characteristics) for you to work on.

Social media analytical technology can help in several ways when monitoring hateful content online. Bear in mind that these software tools are developed for use in richer economies and not developing countries. Keeping the machines and humans' interaction in view, one should be thoughtful of the pros and cons that need to be considered while employing artificial intelligence and machine learning tools. Below are some of the main ones:

Advantages	Disadvantages
User-friendly and practical ✓	✗ Usually, licence is costly with no exclusions for not-for-profit and other organizations
Easy and time-efficient mining of social media feeds for relevant results ✓	✗ Not all social media companies allow or share application programming interfaces (APIs) with third party tools, e.g. Facebook, Instagram.
Easy and time-efficient search of old data on social media and public forums ✓	✗ Social analytical companies can limit the access to old data, make it subject to the package and/or charge extra
Quick and effective way of sorting and managing bulk data ✓	✗ Yet not all data being collected are relevant. Data cleansing is needed
Some tools allow for rapid and sophisticated analysis of results: sentiment and emotional analyses, demographics, etc. ✓	✗ The quality of the search results analysis depends on the effectiveness and the cultural sensitivity of the tool and the data cleansing thoroughness
Auto data visualization using multiple sets of variables, such as timeframe, demographic characteristics, location, sentiments ✓	✗ Sometimes the inbuilt data visualization does not provide what you need. You can easily export the data into other systems and work on visualizations manually
Auto-generated daily, weekly, fortnightly, and monthly reports ✓	✗ Auto-generated reports only contain charts and graphs. Narratives and summaries need to be produced manually
Email alerts can be activated when a certain type of hateful content is high or rising ✓	✗ Email alerts can flood your inbox with spam

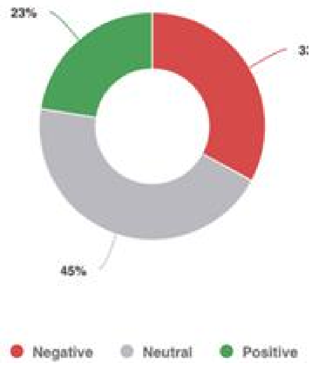
Sentiment and emotions

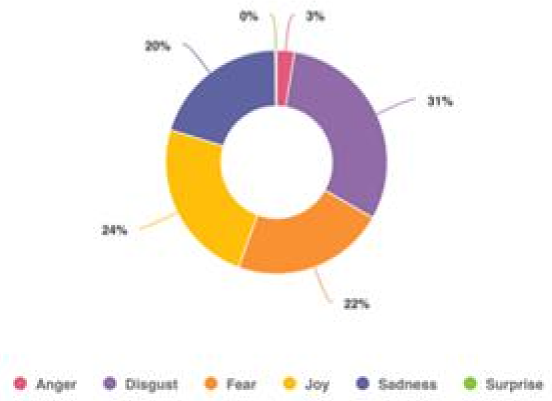
Some tools sort the conversations into neutral, negative, and positive. Monitoring software does this by machine learning about emotions based on emotional signal words close to the key word in a post.

However, **it does not guarantee accuracy.** This sentiment analysis is a hard thing for machines to do – although sometimes easy for humans.

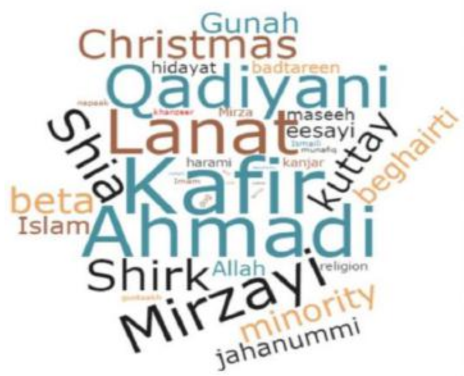
It is wise to go through data with negative connotations and check that the machine has interpreted it right. If your culture includes a lot of sarcasm or irony, this is particularly hard for machines to sort in terms of positive from negative so expect to do a lot of checking.

Similarly, moving beyond simply positive and negative, some software aim to sense specific emotions expressed in a certain contexts, so sorting into anger, fear, disgust, joy, sadness and surprise. This is really helpful in allowing you to understand which types of hateful reactions are linked with viral content or posts that are widely spread.





Themes and topics



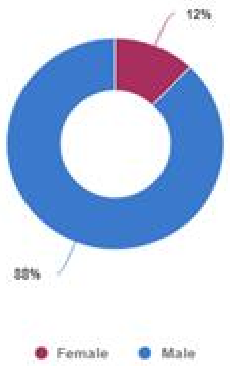
Word clouds sort data based on the most used keywords, hashtags, etc.

Topic wheels categorize data based on the most found keywords in topics and sub-topics.

For those tools offering either topic wheels or word clouds, both are great and easy ways to produce visuals to show people an overview of which hateful terms are trending in a given period.

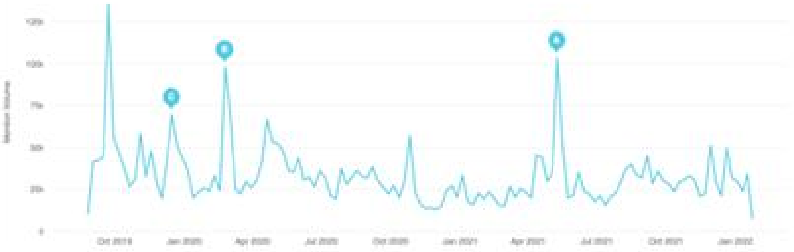
Demographic analysis

Tools which allow for demographic analysis, have the option to visualize information related to gender, profession, and interests, allowing you to see trends on hate-filled content within society.

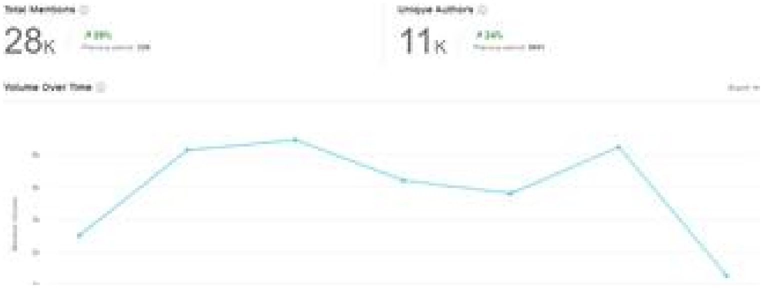


Query timeline and conversation peaks

Query timeline allow you to see patterns over time and conversation peaks which may be linked to real-world events or social media campaigns.



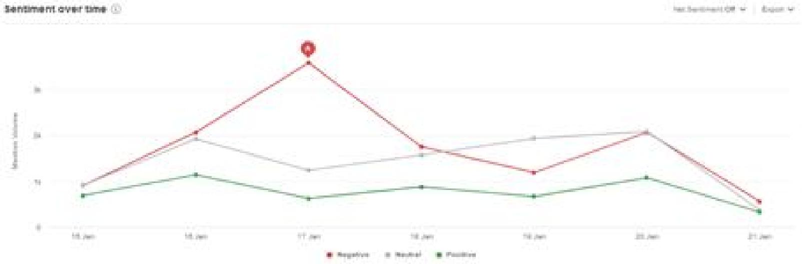
Volume



The volume option (if provided by your tool) displays the total volume of conversations in terms of mentions of keywords and unique authors

Online behaviour

Tools may tell you on which days, weeks and hours, people are more active. This might tell you that use of a particular hateful term occurred after a weekly sermon, during or after working hours.



To get the most out of the above features of a monitoring software tool, on some platforms you may be able to set up customizable dashboards which allow you to tailor the information in much more detail and get into an in-depth analysis of all kinds of hateful expression in your context.

Top social media analysis tools

- **Consumer Research/Brandwatch**, paid
- **CrowdTangle**, free access on application
- **Meltwater**, paid
- **Oracle Social Relationship Management**, paid
- **Socialbakers Suite**, paid
- **Talkwalker**, paid

DISCLAIMER: The guidance provided in this publication concerning the use of any third party social media monitoring tool and analysis of the results thereof represent the opinions and assumptions of the authors. They do not reflect the actual use requirements or restrictions enforced by any of the service providers mentioned above. Your use of each tool will be governed by specific terms and conditions provided by the service provider and you will be solely responsible for your compliance with any agreement entered into with any third party.

2.4 Framing your query

Haroon Baloch

Regardless of the software you decide to use, writing a query is the most important and tricky technical step for digging the right information out of the ocean of online conversations.

Keywords

- **A relevant keyword is the key for an effective query; we need to think of all common, relevant, and frequently used keywords on our theme.**

  For example, for data on hate targeting faith-based groups, we can include the names/titles of all religious groups in the query, such as Hindu, Christian, Shi'a, Ahmadi, etc.

- **It is best NOT to limit our keywords specific to a certain action, incident or event for long-term monitoring, so that we can collect enough data to analyse.**

  For example, a search query only containing keywords such as 'Ashura', 'Christmas', 'Holi', 'temple attack', etc. would not work effectively .

- **A good query must include common hate expressions frequently used by people.**

  A lexicon of hateful terms is helpful.

  See section

  For example, in Pakistan, 'kafir', 'infidel', 'blasphemy', 'wajib-ul-qatal', 'choora', etc.

- **To fetch more, yet relevant data, use different variations of a keyword, including in local languages, such as Urdu or Hindi, by changing your keyboard.**

  For example, Ahmadis in Pakistan are more frequently referred as Qadiyani, which also has other spellings, such as Qadiani, Kadiani, and Kadiyani (قادیانی).

Logical operators

Find out more

Almost all social media monitoring and analytical tools come with common search operators, which are used for broadening and narrowing down the search results. However, only a careful use of these operators can bring out the relevant results.

AND **is used between two keywords**

For example, 'Hindu' AND 'attack' means we are interested in trends involving both of these two keywords.

OR **is used between two keywords when we are unsure if the content containing the desired set of keywords will be available. It broadens the search results**

For example, 'Hindu' OR 'Christian' will return every post that includes either one of them

NOT **is used when we intend to narrow down our search results**

For example, 'Hindu' OR 'Christian' NOT 'Sikh' returns all the results that include either the term 'Hindu' or 'Christian' UNLESS the content also includes 'Sikh'. All posts that contain 'Sikh' will be excluded from the trend even if they include the other terms.

Location and language

After writing a set of keywords of your choice, apply location and language filters. These last two help to get specific results from a specific geographic region, country, district, or city. Similarly you can select only content in one or more particular languages.

2.5 Gathering data using self-reporting

Case study

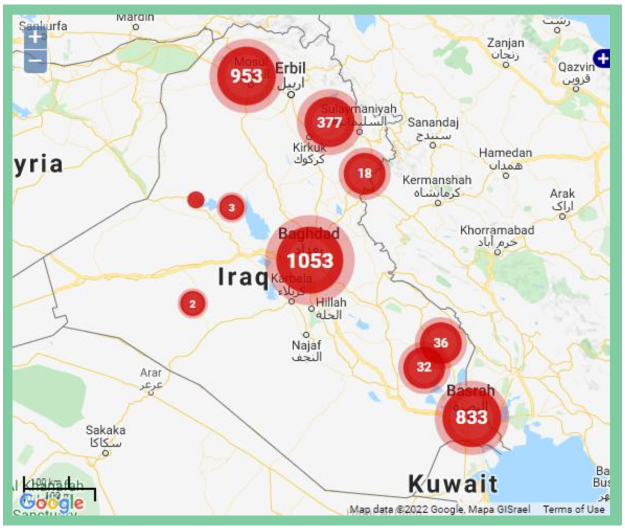
Ceasefire: Monitoring via a secure online platform and crowd-sourcing information

Cecilia Bisogni

As UN rapporteurs and other official international monitors are effectively denied access to a wide range of insecure territories around the world, civilian monitors have become a valuable source of information – in some cases, the only one available – about what is happening on the ground to civilian populations.

Ceasefire Centre for Civilian Rights and MRG worked together to implement a system of civilian-led monitoring of human rights abuses in Iraq, focusing in particular on the rights of vulnerable civilians including vulnerable women, internally displaced persons (IDPs), stateless persons, and ethnic or religious minorities. With the support of Essex University (Department of Computer Engineering) a new online reporting tool was developed which remains available to report human rights violations, as well as to access data and reports.

The system was established after 2014, when ISIS overran much of northern Iraq. It continued to function despite a desperately violent conflict and extreme persecution of almost all minorities in ISIS-controlled areas. Between 2016 and 2020, 3,308 reports were submitted to the system, documenting all types of human rights abuses witnessed or experienced by individuals. The data was aggregated and analysed with regular bulletins issued identifying trends in types of violation, perpetrator and victims. These bulletins published by Ceasefire and MRG generated a great deal of interest in key advocacy capitals and fora including with Iraqi and Kurdish authorities, particularly in relation to minorities and IDPs.



Reported cases by location 2016-2020
Snapshot obtained from the live map on the Ceasefire website

Ultimately, the Ceasefire system allowed continuous and regular data about human rights abuses to be issued to keep the eyes of the world on what was taking place in Iraq, and could equally be used to collect damaging speech data. The crowd-sourced nature of the information did not detract from its authority or credibility as the information was triangulated and supplied by such a high number of individuals reporting from many different discrete locations. Particularly in a setting where any routine data collection by external human rights monitors would have been impossible, the system proved its worth. Positive influences in law-making in Iraq were achieved thanks to the combination of direct advocacy activities and active collaboration with national policymakers both in Erbil and Baghdad in relation to protection of minorities and enforced disappearance legislation, as well as anti-discrimination policies in schools at the provincial level.

10 steps to...

Gathering data using self-reporting

1 **Research** thoroughly the context to understand the nature of the problem you want to address and the main challenges faced by different communities

2 **Choose** a package and develop a set of questions. You will need to avoid asking everything possibly relevant and making it too long and be clear about your priorities. Open-ended questions take a lot of work to analyse, multiple choice take less

3 **Be aware** of personal bias and assumptions. Choose words carefully. Avoid leading and judgemental questions. Where possible, ask key questions twice in different ways to help verify results. Provide as many languages as relevant to the target communities to complete the form

4 **Ask** sensitive questions and the respondent's personal details (if applicable) at the end, to build trust and increase a sense of confidentiality. Only commit to take action on a report if you are sure you will have the resources to deal with these requests

5 **Introduce** your organization and its aims, and say how the information collected will be used. Clearly explain the benefits of the questionnaire for the respondent and assure the of confidentiality

6 **Ensure** that the form is very easy to use. Provide guidance for anything the user may not be familiar with. Define keywords, e.g. discrimination. Pilot it, test it on different devices and using different software for glitches, learn from the feedback and make improvements

7 **Ensure** that you have a safe place to store the data and good security in place, especially if it includes anyone's personal details/contact information

8 **Consider** different ways to get the form filled, e.g. via phone or face-to-face interviews and disseminate it through trusted local organizations and civil society networks. This will build trust over time and allow you to reach people who may not have access to online reporting tools

9 **Publicize**, publicize, publicize (you will need a budget), through social media and media platforms, newspaper or radio adverts, and leaflets

10 **Make** periodic assessments by analysing data and gathering feedback. Publish the results (respecting anonymity) and use the data for advocacy

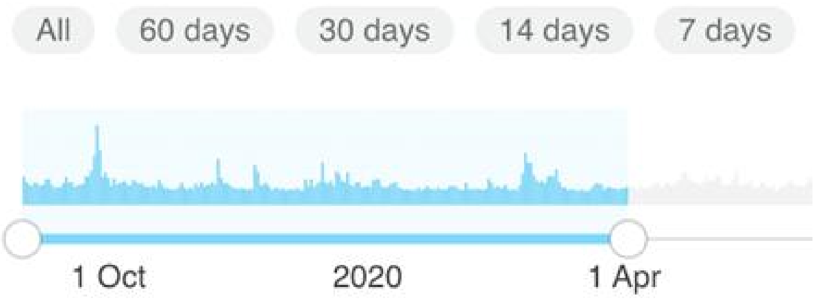
2.6 Analysing results

Haroon Baloch

Analysing search results is the most important step (requiring extra vigilance and care) for documenting and understanding trends. If you use the tools wisely, you may be able to begin to understand where all different kinds of hate come from, what prompts it and how offline and online events interact. For many monitoring software packeges, the following technical aspects will help you in analysing the results:

Use of the timeline

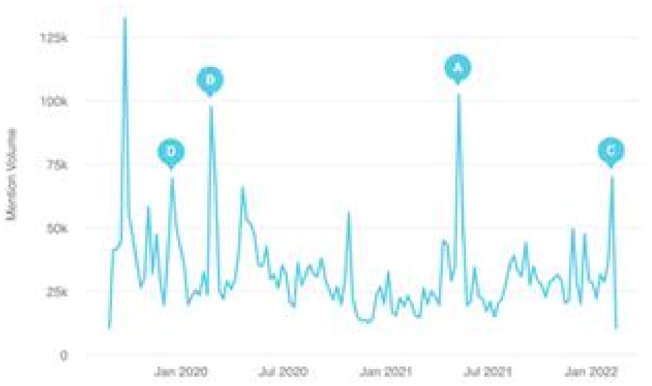
Some tools provide a timeline. You can drag the pointers and select the desired timeframe. As soon as you change the timeframe, the programme will change the results and their visual presentation.



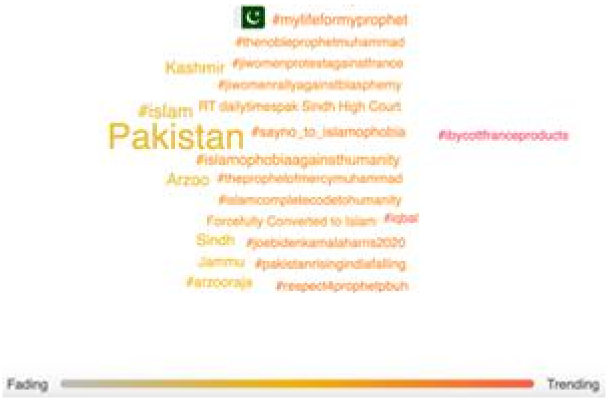
Understanding the peaks

The results are usually presented in the form of peaks on a graph where the x-axis corresponds the timeline and the y-axis corresponds the volume of results. The volume varies from day to day, with clear 'peak days' when the found content is high in comparison to other days.

These peaks indicate that a certain day has generated more results on the query. They can very conveniently be clicked and explored on social media platforms, e.g. Twitter. Often a peak is linked to an announcement, sermon, public debate or media story which helps you understand the genesis of each peak.



Similarly, using trending topics, one can easily understand topics which are trending or fading. The size of the text also conveys the volume of results on a certain topic. The results can also be sorted and analysed further based on disaggregated information using options of demographics, geography, online use, etc.


#mylifeformyprophet
#thenobleprophetmuhammad
Kashmir
#iwomenprotestagainstfrance
#iwomenrallyagainstblasphemy
#islam
RT dailytimespak Sindh High Court
Pakistan
#sayno_to_islamophobia
#iboycottfranceproducts
#islamophobiaagainsthumanity
Arzoo
#theprophetofmercymuhammad
#islamcompletecodetohumanity
Forcefully Converted to Islam
#iqbal
Sindh
#joebidenkamalaharris2020
Jammu
#pakistanrisingindiafalling
#arzooraja
#respect4prophetpbuh
Fading
Trending

Not all the results in a certain peak, or any other option, may be relevant. It is advised to read out the results carefully, and only include the relevant data for reporting. This will require manual work for data cleaning.

Download the relevant data in Comma-Separated Value (CSV) format for manual analysis in Microsoft Excel or Numbers (for Apple).

See more information

2.7 Anticipating hate peaks

What you need to know

You can get ahead of a hate speech spike

In some contexts, a surprising proportion of events triggering hate speech peaks are really very predictable. If you know a spike of hateful content is likely to occur, you can get ready for it, prepare your materials, forewarn your influencers, clear your team's diaries, etc. ☞ See section

From your monitoring, especially once you have done one year or more, you will be aware of certain regular events that will trigger a spike in hate. Be warned, however, that in one country an influx of refugees who are a religious minority may trigger a hate spike targeting that community. In another country, while the arrival of refugees won't trigger any hate, an announcement by the government of a positive measure in favour of minorities will result in a deluge. You need to know and understand your own context!

Below are some situations we have identified thus far. Some are very predictable. Others you might have inside information about.

⚠ Highly predictable

- **Religious holidays and festivals**
- **The opening of a new religious building**
- **The appointment of a new religious leader**
- **Any kind of election**
- **Significant national anniversaries** (independence, victories, invasions, leaders' deaths, which prompt exclusive nationalist and anti-minority rhetoric which can shade into prohibited or restrictable hate speech)

⚠ Somewhat or possibly predictable

- **Any pro-minority government decision** (you might know about it, you might even have been consulted, so even if you are not sure of the timing you can prepare)
- **An event or process in another country where the majority in your country are under threat** (e.g. Israel/Palestine, Kashmir). There may only be a few hours' delay between the international event and it turning nasty domestically so you may need to act fast
- **Any perceived threat to the power of the elite, government** (when they fight to hold power, they may fight dirty)
- **Any major perceived new threat to the economy of your country** (e.g. a trade concession being revoked)
- **An unpopular court decision in favour of a minority individual**

2.8 Monitoring hate during elections

Case study

How Peace-Point Myanmar (PPM) monitored the election campaign period in Myanmar in 2020

Htet Swe

The election campaign period for Myanmar's 8 November 2020 general election began on 7 September 2020. PPM staff monitored the related campaign speeches, platforms and public statements by prominent actors (12 political parties, 22 candidates and 7 media organizations) to document instances of hateful expression and racist and discriminatory disinformation. The project staff also attended some election campaign rallies and events run by political parties.

Based on our experience, the three main issues related to hate relevant in our context that should be researched and/or publicized are: analysis of the types of hate speech; percentage of election campaign hate-filled material; and leading authors and publishers of hate-based papers. We were also interested to understand how audiences interpreted statements, noting the need to raise media literacy awareness to reduce the impact of hateful expression. See section

Social media influencers, comedians, artists, the film industry, religious leaders, and local news organizations all play important roles in spreading hateful misinformation. It was clear that rural areas saw more hate expressed than urban ones, with a particular focus on the north-western area of the country. Whenever a Rohingya candidate competed in the general election seeking to represent the Rakhine state, hate spikes against them appeared. Finally, a lot of hate was being spread systematically online by regular supporters of political parties (many of whom were financially supported by those parties). It was noted that, despite having in place adequate election rules, regulations and laws, the Union Election Commission took no action against hate speech prohibited by Myanmar law during the 2020 election period, even where incontrovertible evidence existed that (prohibited) hate speech was spread by candidates.

In comparison to past election periods, due to Facebook's bans and restrictions, hateful content was particularly common in the comment sections of Facebook posts, news ads and closed Facebook groups. Targets of hate were mainly Rohingya and Muslims. Go to example

Challenges and difficulties

Due to Facebook's bans and restrictions:

- People started using inferences and code words which were more difficult to capture and count as conveying hate (especially using software), even when read within context.
- People began spreading hateful content via TikTok, with which we had less experience in monitoring, and we did not have enough time to do this effectively in the tight election period. See section

Peace-Point Myanmar

Community based non-profit organization, established in 2016 to promote and protect the civic rights of ethnic, religious, and marginalized minorities – the Rohingya.

Peace Point

Regions: Yangon, Rakhine State

Find out more

10 steps to...

Monitoring hate during election periods

1 **Put** together a team of 4–12 individuals who are unbiased (whether by party, ethnicity or religion). If you are planning to monitor in several languages, you will need native speakers or those fluent in each language

2 **Ensure** that your team has a thorough understanding of the background context and a very clear understanding of what prohibited hate speech is in international law and in your country context. You need them in place around 3–6 months before the election

3 **Decide** using objective criteria when you will start monitoring and who you will target or prioritize, e.g. focus on certain newspapers or online sites, certain parties or certain geographical areas. Monitoring during elections can be very time consuming

4 **Deploy** your team to analyse online content, read print newspapers, visit websites, read speeches and policy announcements, follow candidates on Twitter, and even travel the country attending rallies

5 **Have** a team member apply to join likely closed social media groups to monitor content trends, if you suspect prohibited hate speech is appearing there. You can get clues from public discussion or just choose at random and follow relevant groups

6 **Keep** an eye out for campaign promises for minority or excluded communities while scanning for hateful outputs. This can be useful to remind elected candidates of their promises

7 **Capture** evidence of every occurrence of hate or any positive campaign promise. Collate and analyse the data by party, area, speaker type, reach and medium. Link the captured evidence with real-world incident reports

8 **Report** prohibited hate speech in campaigns to local or national election commissions (or international observers) and track whether or not they take action. Also, report to people (citizens) so they can be aware of hate speech and to raise awareness among them

9 **Summarize** your findings, publish a report and disseminate it to all political parties and human rights groups active in your country and internationally

10 **Look after** your people

See section

What you can do and how

Responding, reporting

3.1 Reporting to social media platforms

If you want to report something on social media, then it must be something which goes against their content policies. All social media companies have established rules, norms of conduct and policies which provide guidance on what is and what is not allowed on their platforms.

i You need to be logged in to report any type of content directly.
Reporting is confidential.

Policy	Report	The level of detail	Visual cues	Forms/emails
🐦	Tweets, comments, profiles	You cannot select sub-options after having selected a main category	•••	🔗 Selection of topics
in	Posts, comments, profiles	You can select categories and sub-categories	⋮ (•••) ⚑	🔗 Harassment and threats
f	Profiles, comments posts, events, ads	You cannot select sub-options after having selected a main category	… ! •••	🔗 Generic: access instructions 🔗 Identity theft form
You Tube	Profiles, videos, thumbnails, comments, ads, playlists, links	You can select from a number of options with their own drop-down menu or 'none of these apply' to explain the problem in 80 characters	⋮ ⚑	🔗 Legal complaints '*Give Feedback*' – menu band on the left on the YT page
📷	Posts, comments, profiles	You cannot select sub-options after having selected a main category	(!) Report... ⋮ !	🔗 For people with no account
♪	Videos, messages, comments, sounds, hashtags, profiles	You cannot select sub-options after having selected a main category	⚑	🔗 General form feedback@tiktok.com Legal complaints: legal@tiktok.com

Case study

Reporting to social media companies using escalating channels

A prominent blogger from Karachi, vocal on hate speech issues in Pakistan, who belonged to a religious minority community, became a victim of a concerted hate campaign on Facebook for her bold stance on religious freedom, politics, and fashion. On social media, she had also been actively sharing her modelling pictures.

In 2021, she did a photoshoot with a clothing brand and, after posting pictures on her Instagram and Facebook accounts, she immediately started receiving negative reactions from people including hateful comments breaching Facebook's (now Meta) policies, religious edicts about her clothing and her religion, and abuse.

Soon after, she did a 'live' on Facebook to clarify her stance; however, it backfired. People criticized and insulted her during her live video and warned her to 'face the music' for her public statements. In this hate campaign, some conversations were provocative and classed as incitement to violence. The blogger reported the threats and the incitement to violence on the platforms but did not receive a response. At the same time, her account was hacked. The hacker copied all her photos and videos and started blackmailing her.

Extremely stressed and distressed, by now living in hiding, she contacted one of her friends, who was in contact with a rights-based organization. Together they asked for help to take down the conversations. The organization agreed that the content breached Meta's content policies. As it had previously successfully joined Facebook's Trusted Flagger Programme, the organization escalated the case with the social media company.
They requested the urgent removal of the hate campaign because they feared the potential severe consequences to the blogger´s health and even life.

Trusted Flagger Programmes

What: Community of flaggers with proven record in reporting content which violates the Community Guidelines of the platform frequently and with high degree of accuracy.

Who: Individual users, government agencies, and NGOs with expertise in policy areas.

Benefits: Prioritization of flagged content for review, input into content policy areas

Joining the programme is difficult though.

See section

This resulted in the removal of the campaign against her, in particular the deletion of comments inciting violence. Meta also closed the account of the hacker. This took about 48 hours after the content was flagged by the trusted flagger organization. Meanwhile, the victim was also advised to remain in hiding until the situation had calmed down. If you do not have trusted flagger status and you face a situation like this, try to find someone in your country who does.

Since the campaign and the act of the hacker fell within the ambit of cybercrime laws, the organization encouraged the victim to approach the national cybercrime agency to file a formal complaint.

3.2 Getting prepared for direct challenges

What you need to know

How to get ready to counter hate

Haroon Baloch and Arsalan Ashraf

It may be tempting to respond directly to hateful messages online. A tweet, a Facebook status or a comment can be useful tactics for alternative points of view to gain online visibility. This approach has its risks, though.

You may be personally attacked, trolled or even threatened online by those using hateful terms. More worryingly, your opponents may create fake accounts in your name and post content that can be very damaging or dangerous for you. Some individuals found themselves arrested and charged with offences for posts they did not publish.

Before you start publicly commenting on offensive posts you need to take some basic steps to protect yourself and your accounts.

Strong passwords

Always use strong passwords for your digital accounts – over 14 characters long including special symbols (*,$,#,@), numbers, small and capital letters. Create different passwords for every account to secure them in case one is compromised. Never share your passwords with anyone and if you notice any suspicious activity, change the password immediately and logout from all other devices .

Two-factor authentication

Activate two-factor authentication or 2-step verification on all your social media accounts and email service providers to fortify their safety and avoid data theft. Almost all social media companies including Facebook, Twitter, YouTube, TikTok, Instagram, WhatsApp, Signal and others provide this facility.

Antivirus

Always install updated antivirus software and run it regularly to keep your device(s) safe from malware.

VPN

Download a trusted Virtual Private Network (VPN) and create a dummy or anonymous account so no one can trace any particular comment or reaction back to you. Anonymity is key when it comes to engaging in online campaigns for social causes.

Monitor trends

Always keep an eye on the current trends on social media platforms. Be vigilant for any development in the online sphere that can translate into offline threats and violence. Keep yourself informed and also inform your friends and family.

Phishing attacks

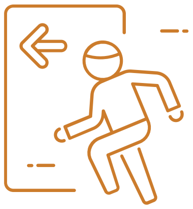
There are no tricks that can earn you millions of dollars overnight, win you a lotto or get you a visa for money. These are some examples of phishing traps hackers use to lure you in and hack your accounts. Never click on any unverified link in emails, websites or messages. Links can contain spyware, trojans, ransomware, keyloggers or other malicious software that can compromise your device.

Be cautious online

Always be cautious and aware of the risks of any backlash while talking about your community or anyone you know, if you belong to a vulnerable group. In case of controversial topics, opt for an anonymous account and exclude personal details: location; specific places of worship, etc. An innocent comment can be misrepresented. A threat could impact not just you, but also others in your community.

10 steps to...

Getting prepared for direct challenges and reacting quickly



1 **Establish** a rota to keep a close eye on developing stories. Often you can spot stories with the potential to generate and propagate hate *Make sure the team know how to flag and report hateful content breaching content policies*

2 **Be aware** of past trends in case you might need to have a response, e.g. linked to religious occasions or news

3 **Don't respond** online if state officials or public office bearers are involved directly or indirectly or engage with people who know you in person *But still report it!*

☞ See section

4 **Use** the digital or manual monitoring tips shared on here to follow closely the hashtags and trends related to an incident and predict its consequences on the community. This should show you where a dangerous amount of public anger exists ☞ See section

5 **Tag** and muster support from prominent progressive personalities, including politicians and celebrities. Try to neutralize the hate narrative with calm, rational or good-humoured speech from people the public trust and respect

6 **Always anonymize** your identity and leave no digital footprints on the internet, in cases where you feel making a response to inflammatory speech is essential. *Ensure your device's GPS locator is off and VPN on*

7 **Avoid** the use of hate messages to counter hatred. It is wrong, counter-productive and escalates the argument and frustration levels, which may increase the risk of violence in the real world

8 **Report** provocative, dangerous and inflammatory online content silently, using relevant reporting mechanisms. Try not to disclose your identity, especially if you represent a vulnerable group

☞ See section

9 **Track** your complaints to record outcomes, where possible, and keep a close eye on the actions taken and their impacts

☞ See section

10 **Look** after the people who are monitoring, responding, and thus, getting exposed to hate. Offer them opportunities to share how it makes them feel ☞ See section

3.3 Getting prepared for emerging challenges

Case study

Countering escalating hate against religious groups on TikTok

Haroon Baloch and Arsalan Ashraf

A worrying aspect, at least for the civil society groups working on the rights of religious minorities in Pakistan, is the way TikTok can be used (or misused) to spew hatred against faith-based groups. The concerns are further exacerbated as TikTok maintains little to no engagement with the rights-based groups in the country.

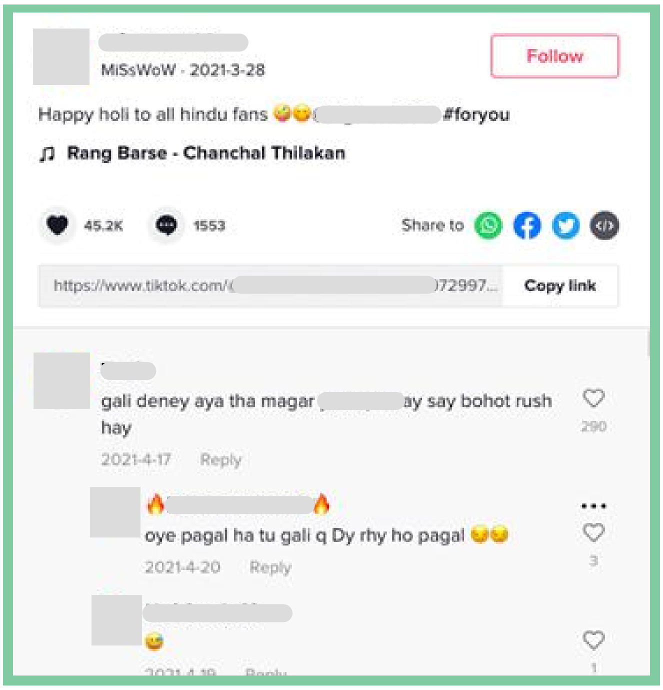
Not only for teenagers

Short video platform providing its users with tools to create 15–60 second bite-sized videos.

1 billion active users – January 2022

19-39 majority age group

Bytes for All has tracked several conversations on TikTok where fanatics can be seen to be targeting Hindus in Pakistan and suggesting that Muslims who exchange Holi greetings with Hindus should move to India.


[redacted]
MiSsWoW · 2021-3-28

Follow

Happy holi to all hindu fans 😜😋[redacted]#foryou

♫ Rang Barse - Chanchal Thilakan

45.2K 1553 Share to

https://www.tiktok.com/[redacted]72997... Copy link

[redacted]
gali deney aya tha magar [redacted]ay say bohot rush hay
2021-4-17 Reply 290

🔥[redacted]🔥
oye pagal ha tu gali q Dy rhy ho pagal 😠😠
2021-4-20 Reply 3

[redacted]
😜
1

A Pakistani TikTok star from Sukkur, Mahjabeen Khan, who is better known as Miss Wow, wished Holi greetings to the Hindu community in Pakistan on 28 March 2021 in a 30-second video, which received more than 1,500 comments. Some of the comments were bitter in taste and people started bullying the content creator by associating her with India (considered by those who stoke division an enemy of Pakistan). Other users can be seen suggesting that the content creator does not belong in Pakistan and should move to India, since she did not greet Muslims for the holy month of Ramadan but did greet Hindus during the Holi festivities.

In spaces which are highly insecure for religious expression, more healthy debates are needed with careful moderation, as unguided debates have more chance of yielding negative implications for already persecuted communities. Ways to achieve this on social media platforms include:

- Take down without delay prohibited speech where vulnerable groups are targeted
- Establish effective content moderation mechanisms in collaboration with rights-based groups, academia and the media who can effectively report all forms of dangerous speech
- Enable anti-hate classifiers on the platform
- Ensure that legitimate and healthy expression is not the victim of moderation.

Rebalancing content

What you can do and how

4.1 Positive messages in the media

Case study

KirkukNow: Countering hate, connecting through audiences

Salam Omer

> *'KirkukNow sheds light on the religious minorities in a professional and unbiased way, played a good role in empowering our voice and gave us better chances to talk to the media. It is important to convey the truth as it is.'*
>
> Christian from al-Qush in Nineveh, Iraq

The tsunami of hate seems to be increasing, with minorities forming the overwhelming majority of victims of online hate.

Between June 2020 and March 2021, KirkukNow's team of journalists produced 50 in-depth reports and video stories, highlighting minority communities' own assessments of their needs and situation and conveying their voice to the public.

KirkukNow's publications have inspired other media outlets to follow: increasing coverage of communities of religious minorities, shedding light on their concerns and promoting minority voices. Based on Google Analytics and Facebook Insights, the 50 in-depth reports and video stories (along with the articles later published as a booklet) reached approximately to 744,932 people on Facebook alone, in addition to the audience who interacted with the articles on other platforms, estimated at tens of thousands.

One of the positive outcomes of KirkukNow's continuous efforts in confronting hatred was winning the trust of religious and ethnic minorities, and enhancing relations between those groups and the media. Once they could 'see themselves' in KirkukNow, they started to proactively reach out with news, events and concerns, seeking coverage in an outlet they could rely on to be unbiased and accurate.

KirkukNow

Independent media outlet, covering developments taking place within or relevant to Iraq's disputed territories.

Provides accurate and impartial news and information to help multi-ethnic and multi-religious communities understand themselves and engage with the surrounding world.

Members of minorities compose a large number of its primary audience.

It aims to join the efforts in place to fight hate speech and fake news targeting these groups.

Team: a large group of trained and sensitive journalists in place on the ground across the country who can be directed towards particular types of stories.

Find out more

Go to booklet: Kaka'is in the time of corona

10 steps to...

Producing positive media content

1 **Research** the working environment. Without prior knowledge of the context, you make yourself vulnerable to many challenges which may later result in problems that are difficult to manage

2 **Recruit** a diverse and politically unbiased team. With a diverse team, you win the hearts of local communities; with an unbiased team, you welcome all different views to the table

3 **Adopt** objective and inclusive reporting. Media coverage should boost the exposure of underprivileged communities through both more and higher quality coverage

4 **Focus** on building your audience and reaching as many people as possible. Also, build relationships with journalists, editors and keep reminding them about your rebalancing content

5 **Highlight** positive developments and success stories of groups at risk of exclusion to balance the flood of negative content published about them

6 **Start** fighting fake news by correcting the imbalance in the media through well-researched and facts-based journalistic articles and verified information

7 **Use** digital media to raise awareness among young people and encourage them to exchange views on a variety of subjects; short stories to introduce religions can also be published on social media

8 **Build** space for open dialogue concerning the commonalities that all communities share; focus on what unites them as well as the benefits of diversity in society to feed into an anti-hate discourse

9 **Set up** a triple coalition between the media, authorities, and CSOs to confront hate-driven narratives where it is safe to do so

10 **Push** governments to look beyond religious freedom laws and to focus instead on implementation

4.2 Positive messages in youth action

Case study

How Bargad supported youth-led Social Action Projects

Iqbal Haider Butt

Young people across Pakistan supported by Bargad, have designed, led, implemented and evaluated 83 Social Action Projects (SAPs) to raise awareness about and tackle hate in society. Bargad already had in place strong outreach programmes for young people in Pakistan before starting this work, and good links with staff and leadership in many colleges and universities. This gave us a very good base to work with. The project followed the 10 steps on the next page.

The young people stunned the project team with their creativity, activism, bravery and insights into how young people consume social media and could be influenced. Even more so because of the Covid-19 lockdown. With universities shut down, the trainees had to look outside their campuses to recruit groups and spread the word. SAPs included poetry, articles and blog writing, debate and poster competitions, signature campaigns, documentaries, interactive and awareness sessions, panel discussions, Facebook/Instagram live sessions, fashion shows, radio programmes, interfaith Christmas celebrations, diversity tours, trips to places of worship, and more.

83

SAPs completed by young trainees in Pakistan

all 4 provinces

2,361 peer youth directly involved

43% women

15% from minority communities *(well above their proportion of the population in Pakistan)*

Many of the SAPs had an outreach way beyond those directly involved and a number of the young people were able to reach large audiences as a direct result of their profile being raised by SAPs, e.g. by appearing on TV or radio programmes to discuss varieties of hate speech and/or religious inclusion.

See examples in context

Testimonial from Sameer Ali Khan

Sameer Ali Khan is a youth volunteer of Bargad who, after attending a CREID training event, mobilized other fellow trainees from Sindh and founded a digital platform named Pakistan Collective. He is now making videos to promote diversity, interfaith harmony and stories of positive peace messaging that bring out the best in people. Sameer and his colleagues were able to turn 120 (until then passive and uninformed) individuals into active and informed social media promoters for peace.

Bargad

Leading Pakistani organization working on youth development since 1998 through a national network of youth volunteers.

The hallmark of the organization is to engage youth in all aspects of the youth-related projects and encourage them to lead and execute their own social action projects on campuses and communities.

BARGAD

Find out more

10 steps to...

Producing positive messages in youth action

1 **Ask** yourself why. What is your motivation? Hateful messaging hits us every day, particularly young people. But remember that hate needs to be understood and felt from the perspective of victims

2 **Build** your team. Discuss and debate what 'hate speech' is, what it does. Build up your own understanding. If in doubt, get help or advice from an expert. *Talking about hate may trigger emotional reactions. Be prepared and sensitive*

3 **Identify** which issues, locations, young people to target. How will you reach them? Why do they need your input more than others and on which issues? Be inclusive as you make these decisions

4 **Prepare** a list of questions on the pressing issues of hate speech, including definition, types, feelings, observations and experiences, positive messaging, and possible responses in terms of gender, religion, geographic area. Test them

5 **Arrange** focus group discussions (FGDs). Bring out personal observations and experiences of hate speech and peace messages to counter it. Compile results. *Remember, young people respond to dialogue, not polemics*

6 **Compile** a training manual using the FGD results. Get feedback. Carry out training. Support trainees to form groups and plan actions. Invite diverse resource persons for increased networking

7 **Create** WhatsApp or Signal groups of diverse peace champions internally and also link them with stakeholders through dialogue events

8 **Allow** groups to design their own actions to address hate. Teams of youth champions can create, implement, record, report and share widely. *The safety of peace champions is the main priority. Ensure it*

9 **Support** young people to reflect on what has worked. What would they do differently? Share this learning across the network and test new ways of action

10 **Make** inspiring case studies. Share within the group and beyond, to encourage others to take on hate

4.3 Positive messages in online campaigns

Case study

How B4A prepared the #IDontForwardHate online campaign

Haroon Baloch and Arsalan Ashraf

#IDontForwardHate 🔗 is an online and offline campaign, which welcomes allies of peace to join hands so that we can promote it together and minimize ongoing violence. The campaign aims to promote peace, dignity and respect for all communities, including minorities. In addition, this campaign also aims to create a more dignified and equal society in which all people can participate regardless of religion, gender, caste, creed, colour and profession.

To promote this message, in late 2020, we approached people from different segments of society including journalists, students, human rights defenders, activists, and government officials. We crafted messages to use offline and online and for different online platform formats. We decided to focus on Twitter and Facebook. Twitter is used by more intellectual individuals in Pakistan, whereas Facebook is the most widely used online platform in general. We succeeded in generating real momentum at the time of the launch with the campaign trending on Twitter (seventh place at national level in Pakistan) on the day.

With this campaign kicked off, citizen journalists from far-flung areas also took active part and encouraged locals to send their photo pledges too. These areas include Tharpakar, Gwadar, Cholistan and Rajanpur, where many religious minority communities and some indigenous groups reside. The areas are remote and rarely benefit from development interventions. Women from these areas, who have not been part of any mainstream activities and who are rarely seen in public events, sent pledges and added their weight. The inclusion of women and people from these areas not only made them part of a national campaign and boosted their confidence but also multiplied the impact of the campaign.

The campaign created a strong sense of social solidarity with the victims of hate, informed those who were less aware of its impacts, and let decision makers know that a large number of people in Pakistan are concerned about and against hateful expression.

Woman supporting the campaign. Credit: B4A

Bytes for All (B4A)

Pakistan-based human rights organization and a research think tank with a focus on Information and Communication Technologies (ICTs).

Promotes the use of technology for sustainable development, democracy and social justice through research for evidence-based policy advocacy, field projects and capacity building of citizens and civil society organizations.

www.bytesforall.pk

🔗 Find out more

10 steps to...

Producing positive messages in online campaigns

1 **Create** a website or a social media account for the campaign. Make sure the website URL or the account name is simple and easy to remember with common spellings. Keep it simple and attractive. *If you have limited experience, try to involve a communications expert*

2 **Envision** a clear target audience. Create a strong and clear message about the campaign. A strong message is one which is simple to understand, touches people and motivates your target group to get involved

3 **Make** it super easy for people to take action. Create a simplified form. Keep the form extremely short and to the point. Test it repeatedly on all formats, platforms and browsers

4 **Reach out** to other organizations or individual influencers to back your campaign. They will help spread the message

5 **Prepare** a strong series of messages in advance to use online and offline. You need diversity and new materials to avoid monotony. Consider whether you can pay for a tool to schedule release of messages

6 **Set up** campaign accounts on all platforms to reach every type of audience. Every platform has a different type and mood of audience. Concentrate on the platforms that suit your target better

7 **Write** separate posts for every platform. They have different dimensions for posts. This will allow people to see the full post while scrolling down their social media timelines/feeds

8 **Encourage** people to create their own messages, but monitor this content carefully. Cross-posting is highly effective if used in a wise way. It helps to keep all platforms up to date and saves time

9 **Get** supporters geared up for the launch day. Be careful your site is not suspended due to 'inauthentic' behaviour. Create a big splash for the long run

10 **Ensure** every campaign contributor feels valued. Thank them. Tell them about the next steps. Publish a report after a month or at the end and share it with relevant stakeholders

4.4 Reaching more people with your content

What you need to know

'Going viral' – how to drive engagement

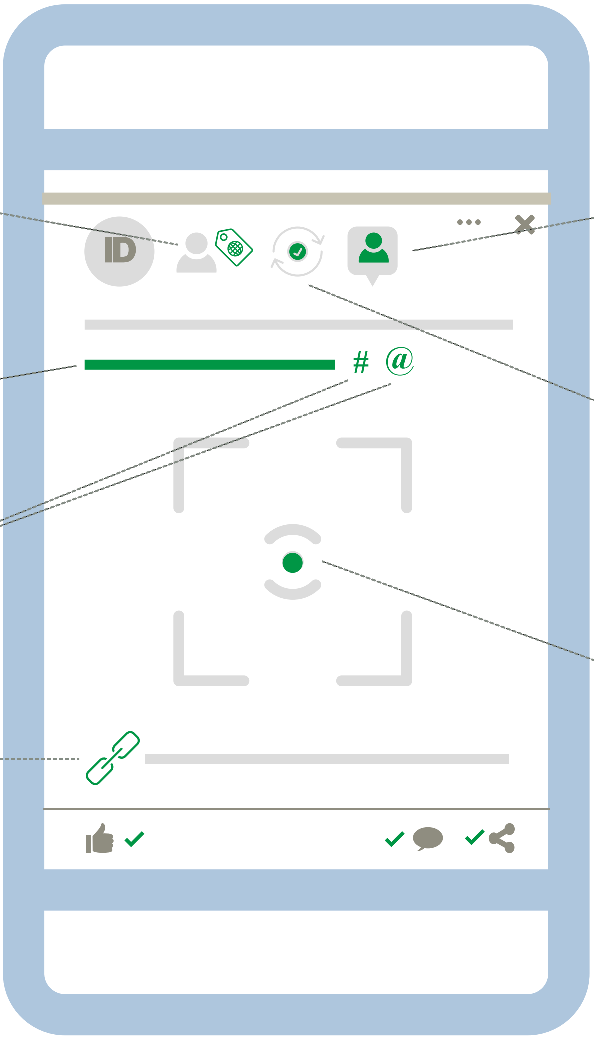
Noah Rosenberg

The main thing is that your post will need to stand out in the 'attention economy' of social media and grab people's attention in the milliseconds it will take them to scroll past it. That will then encourage the algorithm to show it to more people. Below, you will find some ways to increase the chances that your posts will be successful.

What makes some posts go viral?

Luck and **the platform's algorithm** determine which posts truly go 'viral'. You have no control over either of these, but you can influence them.

At the level of design



Tag users with larger followings. This can backfire. Be cautious and only tag users you trust

Keep your text as short as possible. Increases the chances of users reading it

Use relevant (and popular!) tags or hashtags on your post. Note that this is also possible on YouTube

Include a link. Studies show that posts including links get shared more, especially on Twitter

Increase your own follower base. Getting a larger user to share your profile is one good way to do this

Post frequently and at peak times. Check what peak times are for your context – they may differ across platforms and countries

Use multiple kinds of media. For example, photo, video or gifs help make your posts more visually interesting and easy to read

At the level of content and language

Novelty

Information that is new to the reader will be more interesting and memorable

Simplicity

Simplicity in both video and text posts has been found to have a positive effect on engagement. Avoid overly complex sentence structure, wording or video design

Topicality

Users are more likely to engage with posts that are relevant to current events or public debate. Engage topics relevant to you and also popular on social media

Emotion

Psychological studies show that posts containing emotional words, such as 'joy' or 'grief', are more likely to grab people's attention and will be more memorable to them. Try to aim for a specific emotion (e.g. anger, humour, etc.)

Don't overdo it! A post that uses too much emotional language might appear less trustworthy and make users less likely to share it

Call to action

Encourage people to interact more with your post, such as, liking, sharing, or watching a video to increase chances of user engagement

Moral/ emotion

Studies also show that posts containing words that have both an emotional and a moral component, like 'hate', 'betrayal', or 'pride', grab attention more easily

Words and emojis have both an emotional and a moral component, depending on language, cultural and political contexts.

Further reading

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, *15*(4).

Nave, N. N., Shifman, L., & Tenenboim-Weinblatt, K. (2018). Talking it personally: Features of successful political posts on Facebook. *Social Media + Society*, *4*(3).

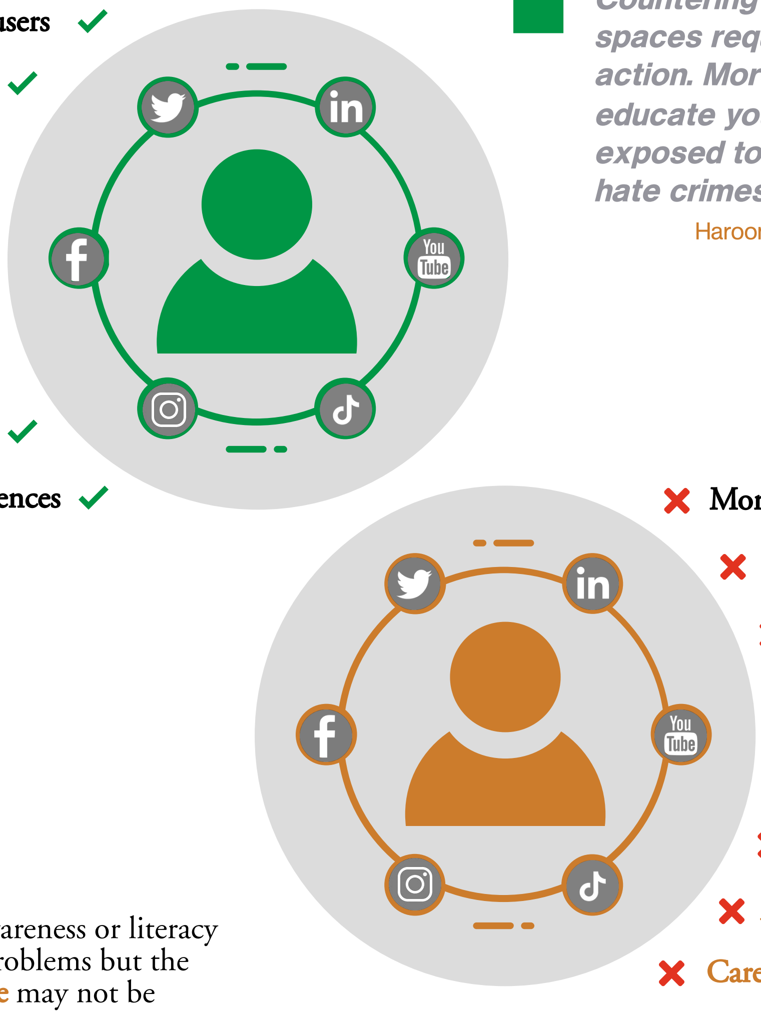
Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26).

4.5 A hate speech aware population

It may now be impossible to eradicate even what should clearly be prohibited hate speech, including cases where a platform tightens its rules and blocks and takes down hateful content. Users may simply move to a different platform with more relaxed policies. New platforms may continuously evolve to exploit this.

But a solution does exist: to persuade the human users of social media to NOT read, watch or share hate-filled content. Provided new generations starting to use social media are educated about hate speech, misinformation and other toxic online behaviours, this could be a sustainable solution.

> *Countering hate in online spaces requires collective action. More important is to educate youth, who are heavily exposed to social media, on hate crimes.’*
>
> Haroon Baloch, Bytes for All, Pakistan



Wiser users ✓

Disbelieve assertions ✓

Think critically ✓

Doubt ✓

Check facts ✓

Assess language ✓

Aware of authorship ✓

Aware of consequences ✓

✗ More naive users

✗ Believe anything

✗ Think passively

✗ Automatically trust

✗ Do not check facts

✗ Do not verify who posts

✗ Actively seek to harm others

✗ Careless about consequences

Improved digital awareness or literacy can address many problems but the two points in orange may not be directly affected (although onward spread of content should be reduced).

MRG´s partner in Myanmar

MRG’s partner in Myanmar designed an outreach session to inform ordinary low-income women about the dangers of believing everything that they see or read online. They planned to convene discussions with women involved in self-help groups and women’s groups. Unfortunately, this work was interrupted by the military coup that took place in January 2021, which made further progress impossible.

Pushing for reforms

What you can do and how

5.1 Influencing Social Media Platforms

What you need to know

Report patterns, analyse, publicize

If you follow even half of the steps in this manual, you will understand the dynamics of hateful messaging and narratives in your context better than most other people. The information gathered during the process could help social media platforms get their systems in order, if they choose to do so. Not, however, in the sense of making them reliant on your free work, but giving them the necessary push to do theirs. Their platforms enable a tsunami of hate and they should be getting better at preventing its impact.

Small grassroots organization
If you can…

Social media corporations
…then they can too

Predict spikes of prohibited and restrictable hate speech

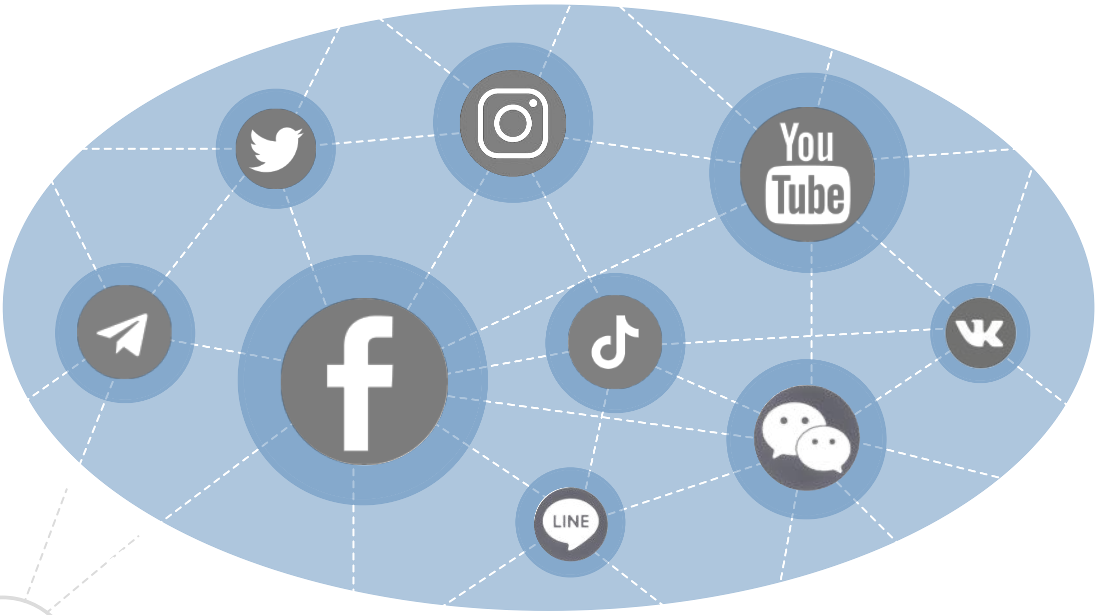
Supply evidence of real-world harm resulting from online hate

Show that hate is not organic but orchestrated, markedly by few individuals

Show that content that breached a platform's policy takes on average 96 hours to be removed, by which time attention has moved on to newer stories

Publicize

One important advantage that you have over social media platforms is that you can understand the entire picture of hate-based expression in your community. Social media companies are rivals, they don't easily share data. But you can show how a story jumped back and forth across platforms over time.



Publish your findings. If it is not safe for you to do so, partner with an international organization who can publish without you being named. Check your facts and evidence exhaustively first. You may be attacked and accused even if there are no mistakes; if there are **any** mistakes, your (otherwise solid) findings will be discounted.

Build relationships with media contacts to help ensure they cover this issue. Look for news outlets and individual articles covering issues affecting your communities well and approach the editors or authors. Cultivate relationships with these journalists and editors.

Engage with social media teams. Become a source of reliable information in your country, so they listen to you and trust you (and act when you tell them that there is a serious problem).

Approach international mechanisms. They are also interested in long-term, large-scale patterns of hate speech in breach of Article 20 of the ICCPR, particularly if you can show how it is generated and enabled, and how the government is effective or ineffective in tackling it.

Draft policy briefs for your government and talk to decision makers you trust when you can.

5.2 Push for improvements in take-down rates, accuracy and timing

What you need to know

Transparency of self-reporting – issues

Noah Rosenberg

Most social media companies now publish regular 'Transparency Reports' reviewing the enforcement of their content policies (including on hateful content). These reports are meant to satisfy the demands of civil society. At a closer look, they leave much to be desired. This section gives an overview of some of the problems regarding the ways in which social media companies report on how they monitor and take down hateful content (whether pre-emptively or in response to reports) and introduces specific means you can use to help build pressure for change.

The review process

Algorithms

There are few details on how algorithms operate or how they are programmed:

Go to Facebook Artificial Intelligence

The share of content removed by artificial intelligence (AI) is increasing. Over 90 per cent of hateful content that is deleted is removed by AI across all platforms. This is a problem because of the lack of transparency:

See AI sentiment analysis

The AIs are not as good as companies claim.

If you see that a post criticizing or satirizing hate is taken down by mistake, keep a record of it as evidence of the problems with AI.

Reviews

There are NO external accuracy reviews on both reported data and removal decisions, making every number reported doubtful:

Go to Facebooke's AI content policy

Go to Facebook Files

The reports only present how much content was removed. Leaked internal documents suggest that Facebook is capturing about 3–5 per cent of the content breaching the community guidelines on hate speech, despite having what many reports say is the best AI in the industry.

Facebook has an Oversight Board that reviews only selective key cases (not content policy appeals) and publishes decisions online:

Go to all Oversight Board decisions

Collect records of how much hateful content you encounter despite its removal. Try to connect with other hate speech monitors to develop an independent estimate of how much hateful content actually gets taken down after being reported.

You might be able to use a relevant decision when meeting with your regional Facebook office and try to influence their operations.

The reported data

There is little to no publicly available data on...

...disaggregation of information by

- **Severity of the violation**
- **Part of a guideline/policy which was violated**

  Even if you manage to get something included in the content policy, you will not get much data on how this specific new section was enforced:
- **Language of the content**
- **Impacted groups**
- **Region/country**

  Experiences of specific groups are not captured.

You will need to monitor the degree of enforcement yourself – whether there are any changes to the previous situation.

...the reach of the violating posts

Many platforms publish no or only unclear counts of views and shares for removed posts.

You may find platforms saying they do publish such data. For example, Facebook may point to its prevalence metric:

Go to Facebook's prevalence metric

However, this is criticized by many observers:

Go to Facebooke's Transparency reporting

Companies make little public effort to connect the spread of violating content to real-world events, such as violence during election periods.

Keep a record of social media posts that you can directly connect to real-world impacts like lynchings. You can use this as evidence against the claims of social media companies that they take down content before it can cause harm.

...the time it takes to remove flagged content

Companies often argue that the take-down time is less relevant than the reach or shares of a post. This does not take into account movement across social media platforms, though. A post may have had no shares on Facebook before it was taken down, but a screenshot of it could be going viral on Twitter.

But be careful: you want to encourage more efficient take downs that still take into account the context of the post, not blanket bans on certain types of expression.

Keep a record of how long it takes between flagging a post and take down. You can use these numbers to build pressure on the companies, as they themselves do not publish this data

Further reading

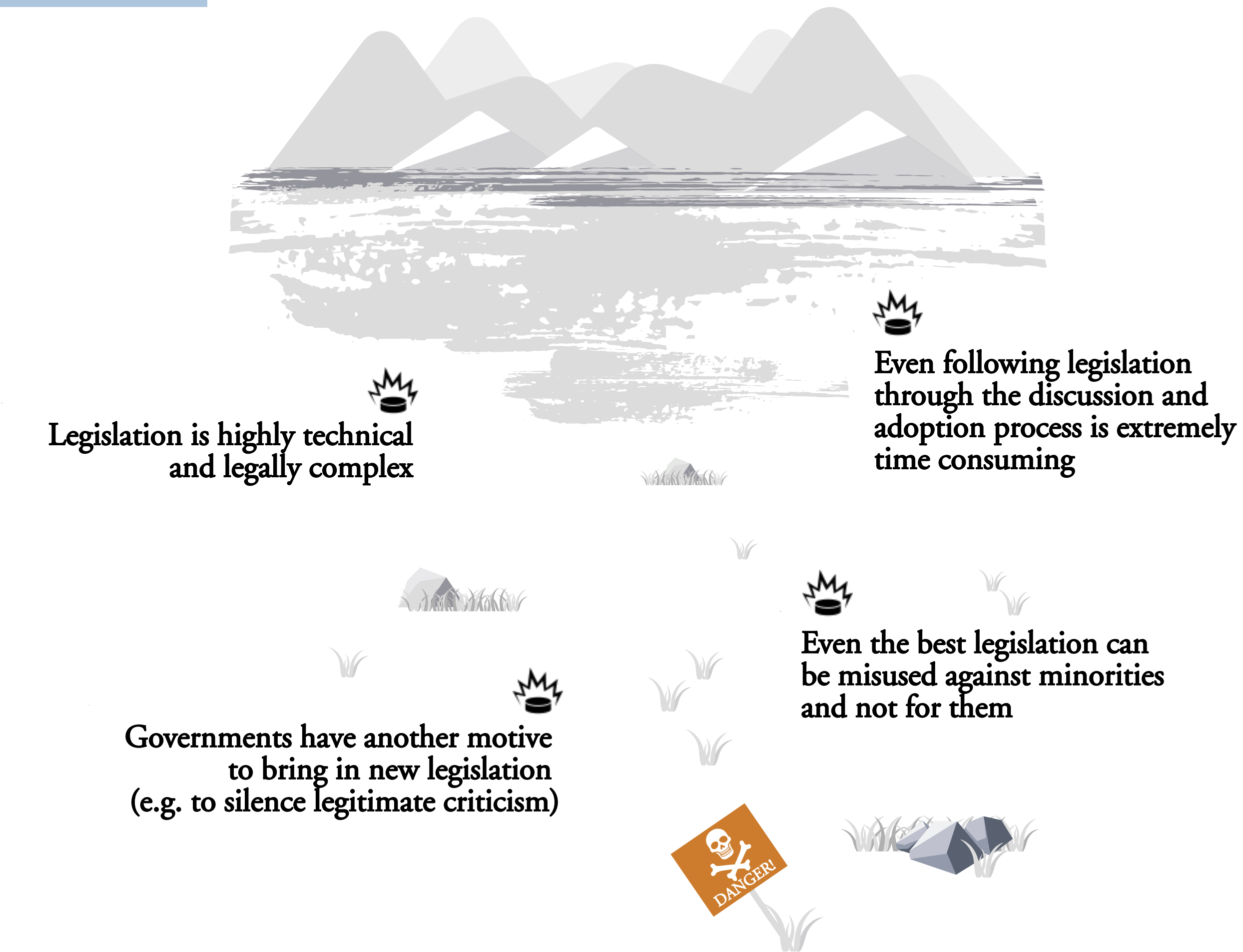
Walsh, E. (2021). Facebook claims it uses AI to identify and remove posts containing hate speech and violence, but the technology doesn't really work, report says. *Business Insider*, 17 October.

Yale Law School (2019). *Report of the Facebook Data Transparency Advisory Group*. The Justice Collaboratory, April.

5.3 Try to influence legislation

What you need to know

For a small organization, proposing hate speech-related legislation might not be the best way to counter prohibited and restrictable hate speech. Why?



Legislation is highly technical and legally complex

Even following legislation through the discussion and adoption process is extremely time consuming

Even the best legislation can be misused against minorities and not for them

Governments have another motive to bring in new legislation (e.g. to silence legitimate criticism)

Help is available

The Office of the UN Special Rapporteur for freedom of opinion and expression might agree to review any legislation passing in your country, especially if you have serious concerns.

Write to them at:
ohchr-freedex@un.org

Join a consortium to oppose new legislation that is not in line with Article 20 of the ICCPR. Many other organizations should already be involved. You should only need to feed in the minority or countering prohibited hate speech perspective.

5.4 Involve international mechanisms

What you need to know

To combat hate speech, the UN has developed a number of tools and policies that can be used by activists

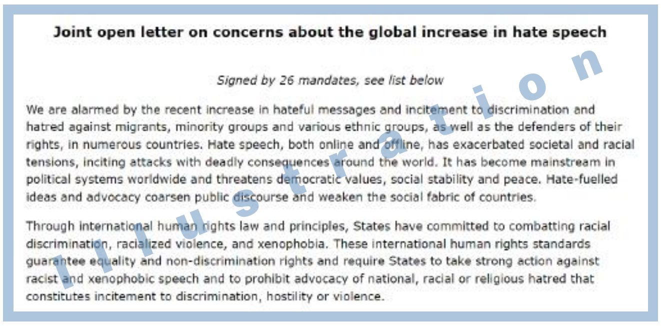
Glenn Payot

Communications or urgent appeals to UN Special Procedures

The UN has a number of specialists who are working on issues and rights that are relevant to hate speech, e.g. Special Rapporteur on freedom of expression, Special Rapporteur on freedom of religion or belief, and Special Rapporteur on minority issues. All of them work on the challenges posed by hate speech, and in September 2019 they wrote a joint letter highlighting their concern about the increase in hate speech and its consequences.

Reasons for documenting most serious cases of hate speech:

- Because it emanates from high-level officials, or from religious leaders
- And/or because it happens repeatedly and/or on a large scale
- And/or because, given the context of the country, it is likely to lead to violence or hate crimes or has done so



Joint open letter on concerns about the global increase in hate speech

Signed by 26 mandates, see list below

We are alarmed by the recent increase in hateful messages and incitement to discrimination and hatred against migrants, minority groups and various ethnic groups, as well as the defenders of their rights, in numerous countries. Hate speech, both online and offline, has exacerbated societal and racial tensions, inciting attacks with deadly consequences around the world. It has become mainstream in political systems worldwide and threatens democratic values, social stability and peace. Hate-fuelled ideas and advocacy coarsen public discourse and weaken the social fabric of countries.

Through international human rights law and principles, States have committed to combatting racial discrimination, racialized violence, and xenophobia. These international human rights standards guarantee equality and non-discrimination rights and require States to take strong action against racist and xenophobic speech and to prohibit advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence.

Access the joint letter

You can share the information and evidence you have, concerning very serious individual instances of hate speech or patterns of prohibited hate speech. For an idea of how to decide what 'most serious' means see the box to the left.

Submit information/evidence

This will invite Special Rapporteurs to react through a letter that will be sent to the government.

Note that Special Rapporteurs receive a high number of communications and are only likely to take action on cases of prohibited hate speech of a particular gravity. The indicators above can help you assess whether it is appropriate to request this.

Engaging with the UN Strategy and Plan of Action on Hate Speech

The UN Office on Genocide Prevention and the Responsibility to Protect is responsible for coordinating UN efforts to prevent, address and combat hate speech through the Strategy and Plan of Action on Hate Speech.

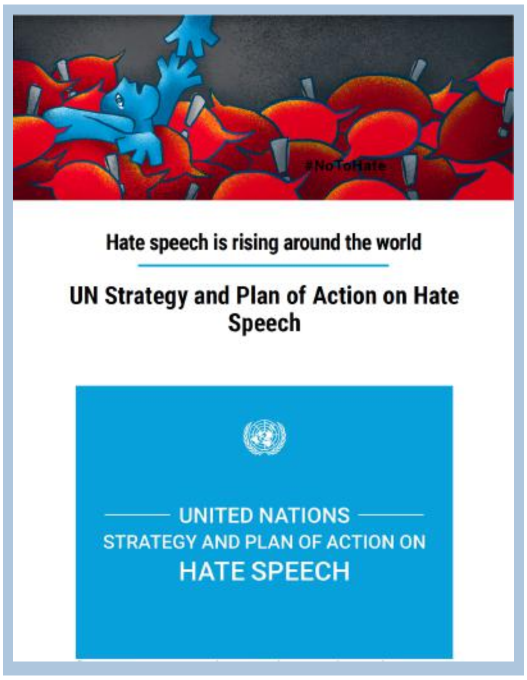
UN country teams (which are the offices coordinating all the UN agencies and programmes in a given country, led by a UN Resident Coordinator) as well as peacekeeping and political missions have a responsibility to implement this strategy.

The UN has developed specific guidelines for them to take part, but in practice some UN country teams are more active on countering hate than others – often depending on the shape and size of their presence and local considerations.

The strategy and guidelines indicate that offices at both levels should be receptive to reporting on different forms of hate speech from civil society actors.

When you have documented serious cases of hate speech you can:

- Contact the UN Office on Genocide Prevention and Responsibility to Protect. Their intervention may involve engaging with you and the case directly, or directing you towards the relevant UN actors in-country. **Contact email:** osapg@un.org
- Contact the relevant UN country team/Resident Coordinator (see below).



Hate speech is rising around the world

UN Strategy and Plan of Action on Hate Speech

UNITED NATIONS
STRATEGY AND PLAN OF ACTION ON
HATE SPEECH

Access the guidelines

The information about the case should include as much of the list below as possible:

- The identity of the emitter(s) of the prohibited hate speech
- The content of the speech (what was said and in what context)
- The medium through which this speech was made (social media, printed press, public discourse, speech in religious context) and information about the audience (followers on social media, number of retweets…)
- The identity of the target(s) and their specific vulnerability
- The context in which this speech was made and the reasons why this speech might be conducive to violence, hate crimes or even atrocity crimes against a particular target

Contact UN country team or Resident Coordinator

To find out the contact details of your UN country team Resident Coordinator:

Copy the link below into a new tab

[https://COUNTRY.un.org/en/contact-us]

Replace the word COUNTRY in the link with the name of your country in English

Examples

https://pakistan.un.org/en/contact-us

https://myanmar.un.org/en/contact-us

If you don't receive a response from any or all of these efforts, don't give up straight away: write back to them about your concerns about 1–2 weeks after you first wrote. Try to identify one person you can speak to and keep sending them examples and evidence of very serious hate speech that you discover and its impacts on communities. They will log what they are sent and the more they receive, the more they will realize how serious a problem this is.

This will also help you to establish a positive working relationship with those in charge of leading on the UN Strategy and Plan of Action on Hate Speech.

What you can't do yet
and why not

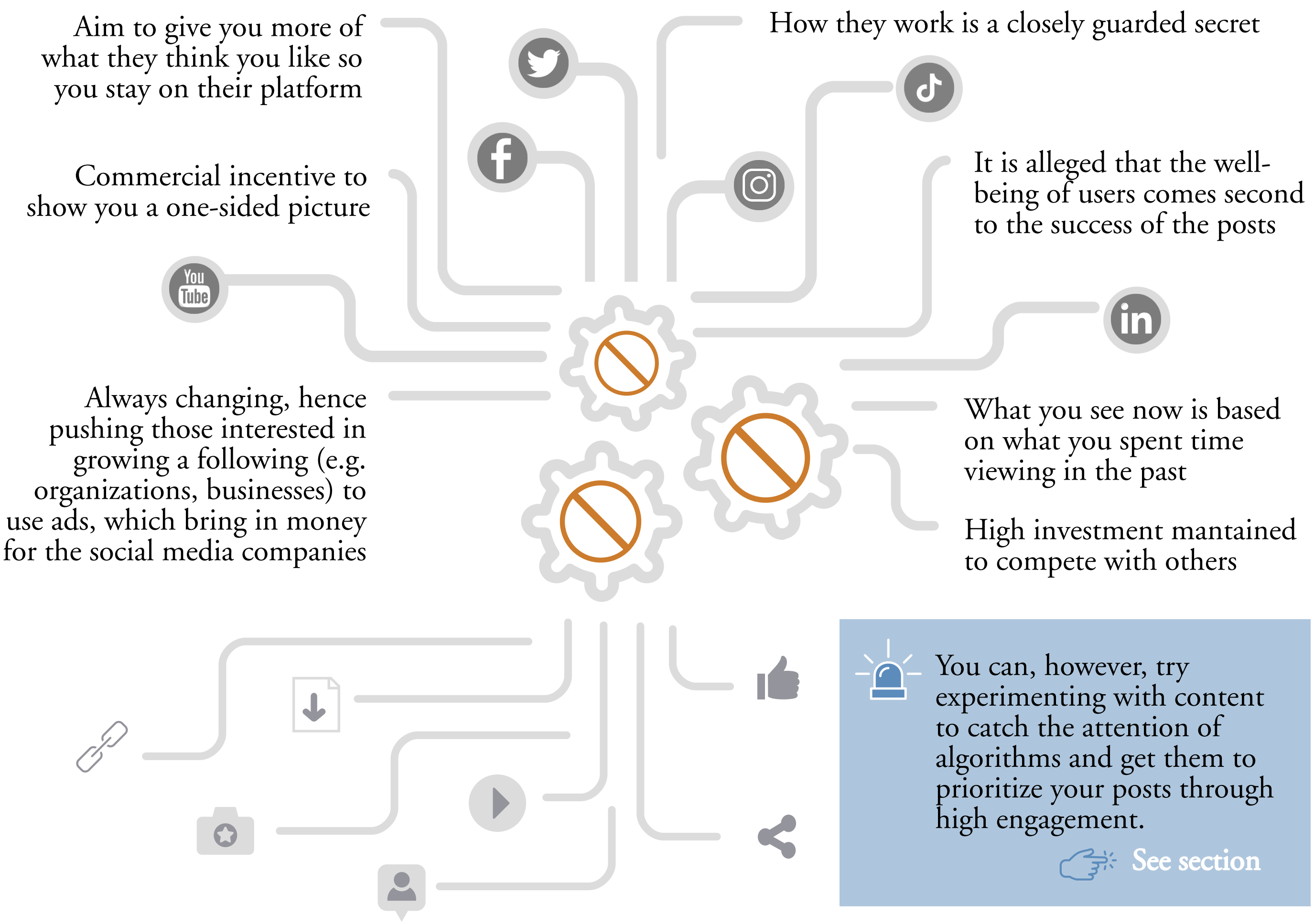
6.1 Influence algorithms

What you need to know

A social media platform's success depends on its algorithms

Algorithms are tools that social media companies use to sort and display content by what they believe to be the most 'relevant' to you, the user, and with the highest likelihood you will actually see it. Why was this needed? In the early days of social media, the newest post would always be on top of a user's feed. As these platforms grew to reach milions of posts, shares, comments, status updates per day, the need to sort content for each user became essential. However, besides improving the user experience, algorithms also gave control to social media companies over what content to show users, based on their behaviour.

You cannot influence social media algorithms; some of the reasons are shown below.



Aim to give you more of what they think you like so you stay on their platform

Commercial incentive to show you a one-sided picture

Always changing, hence pushing those interested in growing a following (e.g. organizations, businesses) to use ads, which bring in money for the social media companies

How they work is a closely guarded secret

It is alleged that the well-being of users comes second to the success of the posts

What you see now is based on what you spent time viewing in the past

High investment mantained to compete with others

You can, however, try experimenting with content to catch the attention of algorithms and get them to prioritize your posts through high engagement.

See section

6.2 Influence content policies

Above we covered how to monitor hateful expresssion and how to report hateful content that breach the content policies of a platform they appear on. Here we turn our attention to what you can do where you have identified damaging speech affecting a minority community that doesn't appear to fall within the content policy of the platform.

Content policies

It is always much quicker, easier and more effective if you could argue that the community in question is covered by the policy! Even if your community is not explicitly mentioned, they may have characteristics that would be covered by a policy.

To take a concrete example

If you are concerned about hateful content against a group of people who are stateless. None of the content policies today explicitly mention statelessness as a factor which would trigger removal of content if discriminatory or violence-inciting material was uploaded. Yes, you can try to get statelessness added to the policy (see below) but in reality a quicker, more effective, immediately available option is to look again at your group of stateless people – perhaps they are racially, ethnically, religiously or linguistically different from others.

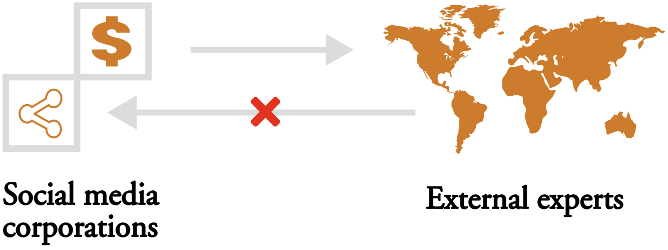
But ultimately you may feel that it is very important to get the platforms to include your issue or your community upfront. Can you influence their policies? It depends...

Policy development teams

An efficient way would be to join their policy development teams, which generally consist of internal and external experts. The latter may include industry and policy experts, academics, human rights organizations or activists.

But you can't apply to join such teams!

They need to approach you ...



High engagement

There is no way to guarantee that they will approach you, but the more active you are on their platform, reporting content and following up, the more useful you are to them, the more likely they are to listen to you or to approach you.

- Report as many instances of hateful content breaching their content policies as possible
- Reach out to, liaise with and prove your usefulness to the platform's team in your country
- Become a member of any trusted flagger programmes

Go to YouTube's Trusted Flagger Programme

If your identified gap in a content policy is urgent and this process is going to be too slow:

- Together with other CSOs in your country write a joint letter to the company asking for change in guidelines
- Try to identify experts who are already consulted by the platform (e.g. MRG for Facebook and Twitter)
- Consider putting your request/letter into the public domain and getting media coverage of the gap and its severe consequences for the community. Companies are sensitive to adverse media coverage and this may trigger more serious action at a higher level.

Conclusion

Look after your people

Scattered throughout this toolkit, you will have seen references to the need to look after the people involved in your effort. But this is so important, we are devoting this page to make sure that the point receives the emphasis it deserves.

Staff or volunteers

regardless whether they are minority or majority...

...may well be disturbed by what they are seeing and hearing when discussing the impact of hateful messages on people.

...may well be disturbed by the content that they are reviewing when monitoring hate or collecting material for lexicons.

- Offer them opportunities to talk about their feelings. For volunteers, this should be built into training or briefings. For staff, this can be as a team, or it can be one-to-one with a trusted, skilled and qualified person.
- Persevere in looking for the right person. It may be hard to find this person in your country, particularly if your issues are very sensitive.
- Remember to include the cost for supporting your staff in your project budget. Defend the budget, when people want to spend it on something else, which they consider is more important. Keeping your people safe and ensuring their wellbeing has to be a priority.
- Make sure staff are trained on maintaining security online and offline. The highest security systems are only as strong as the weakest person operating or implementing them.
- Bear in mind that volunteers may have a lower level of knowledge about security and wellbeing and may be more reluctant to ask for help. They may also be more likely to take risks that put them in danger.
- Have a contingency plan in place to deal with a major security crisis or incident in the most security sensitive settings, e.g. a threat to your staff, a raid on your office, a major hack of your IT systems.

Final word

I believe that irrespective of the difference in colour, race, culture and religion we are creatures of one “Almighty ruler”. I also think that there must be official platforms where religious and social discrimination is reported and the State can remedy against such exclusion.’

Hindu student at Punjab University, Pakistan

Social Action Projects in Kapur district, Punjab, Pakistan Courtesy of Bargad

Kakai minority women´s rights activist, Kirkuk, Iraq. Courtesy of KirkukNow

And also I can say that I criticized and interacted with other different people in negative way, and I thought that they are not as good as us. So, that was hate speech on the different people. However, I came to respect other different people and feel I am increased in empathy on them. Now, I analyse news and voice whether it is real or hate speech.’

Participant
Counter hate speech workshop, Sittwe, Myanmar

My participation with co-workers helped me to identify the environments near me in which there are many cases of hate speech and how to deal with it in a responsible way.’

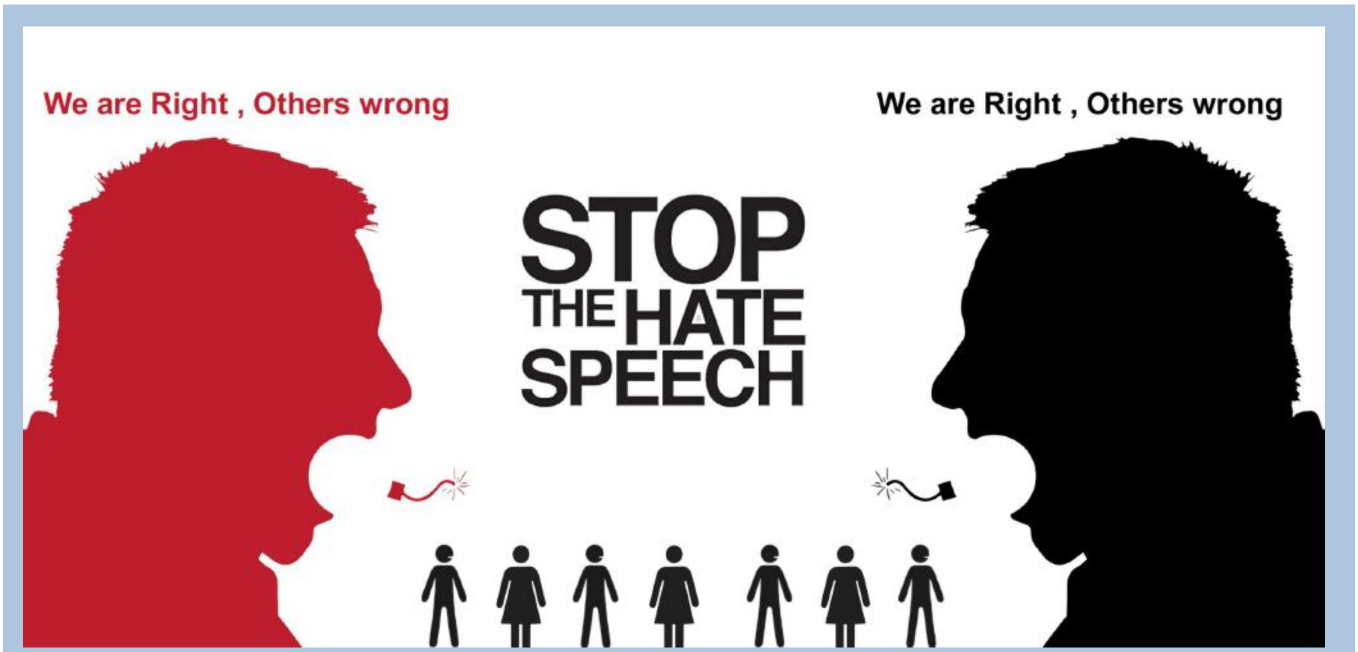
Participant
Social Media Monitoring Workshop, Erbil, Iraq

CREID FGD participants, speaking during the launch of an Interfaith Harmony Project, Pakistan. Courtesy of Bargad

This box represents the ongoing efforts of activists working in Myanmar to counter hate, about which we can’t publish photos since the military coup in January 2021 impacted severely on their security.

Messages and awareness around hate speech should be conveyed via fun, entertainment and fashion because young people are more attracted towards entertainment.’

Youth activist, Peshawar, Pakistan



Youth-led SAP supported by Bargad. Credit: Ahmed Zaeem, Ahmed Zafar, Hassan Shirazi. Poster by Waseem Akhtar

> *More needs to be done to counter such rhetorical discourses of hate and toxic influence. More needs to be done to encourage a language of dialogue, tolerance and finding of mutual understanding of our humanity regardless of religious differences and or ethnic, racial and sectarian differences. Focus should be on what unites Iraq's rich and diverse community rather than on what divides them according to identity politics.'*

Pshtiwan Faraj
IMOK, Iraq

> *At the same time, internet companies must not allow their content ranking algorithms to flare up hate narratives. The states also have responsibility of engaging with the instigators of hate in offline spaces.'*

Haroon Baloch
Bytes for All, Pakistan

> ***Hate speech is a menace to democratic values, social stability and peace. As a matter of principle, the United Nations must confront hate speech at every turn. Silence can signal indifference to bigotry and intolerance, even as a situation escalates and the vulnerable become victims.'***

UN Secretary-General António Guterres
(May 2019), Foreword of the UN Strategy
and Plan of Action on Hate Speech

> *Ask any parent of a young child – to be human is to ask questions. Hate speech is now endemic. No cure or vaccine is available. No software, algorithm or law will ultimately solve this. The solution to hate speech is not to ban or control it. Rather it is to empower consumers to challenge, to doubt, to check and ultimately to ask questions.'*

Claire Thomas
Deputy Director, MRG

Useful links

Introduction

- ‘When the blood starts’: Spike in Ahmadi persecution in Pakistan, *Al Jazeera*, July 2021
- Pakistan’s social media is overflowing with hate speech against Ahmadis, *The Diplomat*, July 2021
- Social media brings both hope and fear for religious minorities in Pakistan, *The Friday Times*, October 2021

About hate speech – What you need to know

- Status of Ratification Interactive Dashboard, OHCHR Indicators, United Nations
- The Rabat Plan of Action, United Nations
- International Convention on the Elimination of All Forms of Racial Discrimination, United Nations
- Convention on the Prevention and Punishment of the Crime of Genocide, United Nations
- African (Banjul) Charter of Human and Peoples’ Rights, African Union
- Declaration of Principles on Freedom of Expression in Africa, African Commission on Human and Peoples’ Rights
- Committee of Ministers Recommendation CM/Rec (1997) 20 on ‘hate speech’, Council of Europe
- American Convention on Human Rights, The Inter-American Specialized Conference on Human Rights
- The Cairo Declaration of the Organisation of Islamic Cooperation on Human Rights
- Report of the Special Rapporteur on minority issues, Human Rights Council, 3 March 2021

Understanding, monitoring – What you can do and how

- National Commission for Justice and Peace, Pakistan, official website
- National Commission for Justice and Peace, Pakistan, contact page
- Independent Media Organisation in Kurdistan, Iraq, Facebook page
- How to use Boolean Search Operators for Social Media Monitoring (and Why You Want to), *Social Media Today*, February 2019
- Ceasefire online reporting tool, Iraq
- Which social media monitoring tools allow export of data for further analysis?, *Quora*
- Policy Briefing 2020, Arakan Front Party, Myanmar, Facebook page
- Peace-Point Myanmar, Myanmar, official website

Responding, reporting – What you can do and how

- Twitter rules and policies on safety and cyber crime, including hateful conduct policies
- Reporting forms, Twitter Help Centre
- LinkedIn Professional community policies
- Reporting Harassment or a Safety Concern, LinkedIn
- Meta Hate Speech policy
- Reporting inappropriate or abusive behaviour (e.g. nudity, hate speech, threats), Facebook Help Centre
- Reporting a fraudulent account in case of identity theft form, Facebook Help Centre

- Google's YouTube hate speech policy
- Legal complaints form, YouTube Help, Google
- Reporting a breach against community guidelines form, *Instagram Help Centre*
- TikTok's Community guidelines, including on hate speech
- TikTok feedback form

Rebalancing content, positive messages – What you can do and how

- KirkukNow, Iraq, official website
- Kaka'is in the times of corona, *KirkukNow*, 2021 (English version, but also available in Arabic)
- Bargad, Pakistan, official website
- I don't forward hate, anti hate speech online campaign in Pakistan, Bytes for All
- Bytes for All, Pakistan, official website
- Talking It Personally: Features of Successful Political Posts on Facebook, academic article
- The MAD Model of Moral Contagion: The Role of Motivation, Attention, and Design in the Spread of Moralized Content Online, academic article
- Out-group animosity drives engagement on social media, academic article

Influencing social media platforms – What you can do and how

- Facebook Artificial Intelligence report, *Business Insider*, October 2021
- Facebook's A.I. content policy, *Fortune*, November 2020
- The Facebook files, *The Wall Street Journal*, October 2021
- Facebook Oversight Board – overview of all decisions
- Community Standards Enforcement Report on Prevalence of Hate speech, Transparency Centre, Meta
- Facebook's Transparency Report, *Anti-Defamation League*, November 2020
- External Review of Facebook's Transparency Report, *Yale Law*, May 2019
- Report on the effectiveness of Facebook's A.I. to remove hate speech, *Business Insider*, October 2021
- Joint open letter on concerns about the global increase in hate speech, United Nations, September 2019
- OHCHR Submission of information to the Special Procedures
- United Nations Strategy and Plan of Action on Hate Speech: Detailed Guidance on Implementation for United Nations Field Presences, United Nations, September 2020

What you can't do (yet) and why not

- YouTube Trusted Flagger Program