Evidence of Divisive Behavior on Social Media

a framework for mediators and peacebuilders

This manual was produced for internal use by Build Up, and is shared publicly as supporting material for mediators and peacebuilders engaging in a social media listening process. For questions about the framework or to request support, please contact team@howtobuildup.org

One of the main reasons mediators and peacebuilders are interested in social media analysis is to understand whether there is evidence of divisive behavior online that is affecting the prospects for peace.

Mediators need this evidence of relevant, divisive behavior online in order to determine what aspects of social media are "mediatable", for example whether a social media code of conduct outlining principles that conflict parties sign up to is appropriate, or whether a clause on the conduct of conflict parties on social media should be included in other agreements being negotiated (e.g. ceasefire agreements). Mediation teams may also need this evidence in order to monitor any agreed code of conduct or clause on social media in an agreement, or to take proactive measures to protect a peace process from social media disruption.

Peacebuilders need this evidence of relevant, divisive behavior online in order to determine what aspects of social media need to be addressed through dialogue or trust-building, for example whether a dialogue with religious leaders is needed to address divisive discourse shared on social media, or whether a narrative change campaign is needed to address intergroup tensions expressed online.

Peacebuilders and mediators may also be interested in social media listening for other reasons – e.g. to add to a conflict analysis, to monitor the impact of initiatives, to understand "talk about the talks", etc. While recognising this, our focus in this framework is what we identify as the most direct reason mediators and peacebuilders are interested in social media analysis: to understand whether there is evidence of divisive behavior online that is affecting the prospects for peace.

With this in mind, and in order to guide mediation teams and peacebuilding organizations in their approach to social media analysis, we offer a framework that first identifies the categories of online divisive behavior that are relevant to mediators and peacebuilders, and then identifies how these behaviors manifest on social media. It also offers guidance on how to gather evidence that the behavior is pervasive / having an impact offline, and how to gather evidence on attribution, including attribution to networks of users.

1. Categories of online divisive behavior

We are interested in **two broad categories of online divisive behavior**:

1. **Deliberate tactics** that some people use to **harass or manipulate** people on social media. These are things that some people do on social media. These tactics are often amplified by algorithms, but they have an attributable source.
2. **Contextual changes** that result from the amplification of harassment and manipulation on social media. These are things that happen to people on social media. These contextual changes are a network effect that results from the interaction of platform design with human psychology – it cannot be attributed but it does need to be understood.

Deliberate tactics and contextual changes interact. For example, disinformation is a deliberate tactic employed by actors wishing to sow division through the production of false or misleading information. Misinformation happens when people unwittingly spread disinformation, without the intention to deceive. Misinformation spreads when people's interests are so polarized that they will believe a piece of information largely based on who shares it: the virality of disinformation is a contextual change connected to changes in what people are interested in. Similarly, coordinated harassment is a deliberate tactic employed by actors wishing to sow division by discrediting an individual or group. This tactical discrediting makes it easier for 'clapbacks' and outrageous claims to go viral: when groups are discredited, the norms according to which people behave on social media change.

Despite this interaction, **we can think of distinct strategies to address each**:

- **Deliberate tactics can be more easily attributed**, and addressing them is about asking people / groups to stop doing something on social media. In this sense, it is more about an agreement to a "code of conduct" or a social media "ceasefire"; addressing deliberate tactics may be of more interest to people acting as **mediators**.
- **Contextual changes are not easily attributable** and may require more proactive strategies to reverse pervasive behaviors or beliefs, such as narrative change campaigns or dialogue-based initiatives. Addressing contextual changes may be of more interest to people working on broader **social cohesion or peacebuilding programming**.

Finally, there are many deliberate tactics and contextual changes that are not relevant to mediators and peacebuilding because they don't affect the prospects of peace. In this framework, **we define divisive behaviors on social media as those that either impact (deliberate tactics) or signal (contextual changes) affective polarization.** Polarization is generally bifurcated into an issue-based or relationship-based analysis. Issue-based polarization focuses on the ideological distance between parties on policy areas.

- **Issue-based polarization** is connected to constructive conflict, the kind that allows for differences of opinion to co-exist in society, and for democratic deliberation that can lead to important transformations in society. A relationship or identity-based polarization is more precisely referred to as affective polarization, meaning the increasing dislike, distrust, and animosity towards those from other parties or groups.
- **Affective polarization** is a dynamic process intertwined with conflict escalation, by which a self-reinforcing spiral cooperates to separate ideologies or identity groups into increasingly distanced and aggregated adversaries.

In other words, where issue-based polarization leads to constructive conflict necessary for a peaceful society, affective polarization leads to destructive conflict that can affect the prospects for peace. Deliberate tactics impact affective polarization; contextual changes are signals of affective polarization. This is why we define divisive behaviors on social media as those which impact or signal affective polarization, and not differences of opinion that impact or signal issue-based polarization.

2. Evidence for deliberate tactics to create division

Conflict actors often use deliberate tactics to create division by **harassing or manipulating** people on social media. These tactics are often amplified by algorithms, but they have an attributable source and a network working (paid or unpaid, identifiable or not) to amplify. One way of addressing them is about asking people / groups to stop doing something on social media. In this sense, it is more about an agreement to a "code of conduct" or a social media "ceasefire"; addressing deliberate tactics may be of more interest to people acting as mediators. Many of these tactics will also violate the terms of service of platforms – specifically if the content incites hatred, harm or violence and / or if it is amplified via a paid network of accounts (automated or human) – and can be reported to platforms via trusted partner or flagger programs.

Deliberate tactic	**Types of behaviors**	**How to find it**
Production of harmful content	Content that explicitly calls for violence or harm towards a group	Search for phrases associated with calls to violence in combination with phrases associated with a group

	Hate speech content that dehumanizes, deindividuates or vilifies a person or group	Search for hate speech terms and phrases
Coordinated harassment of individuals or institutions	Flooding of targeted account with similar narrative or set of phrases Accounts functioning as hubs directing their followers to harass specific targets Accounts providing weaponized talking points to invite others to harass – often personal defamation and / or disinformation about key social, historical or political facts Doxxing of individual accounts to enable offline harassment	Search for phrases that are repeated exactly, in a short period of time, especially in comment threads or reply threads (often on just a handful of posts) Search for links to known pieces of disinformation shared repeatedly
Coordinated harassment of identity groups	Flooding of hashtags with similar narrative or set of phrases Accounts functioning as hubs directing their followers to harass specific targets Accounts providing weaponized talking points to invite others to harass – often hate speech, association with certain traits or fear speech	Search for phrases that are repeated exactly, in a short period of time, especially in combination with a hashtag
Inflation of positions along a dividing line	Flooding of hashtags with similar narrative or set of phrases Network of accounts with few followers (likely to be bots or paid trolls) Impersonation of accounts	Search for phrases that are repeated exactly, in a short period of time, especially in combination with a hashtag Search ad library for the same phrases, and examine targeting

	Use of targeted ads to promote messaging Systematically flagging content critical of a position or narrative	

Identifying whether the behavior is pervasive / having an impact

- Measure how many people it reaches, by looking at followers and shares
- Is it reaching people outside the follower network of the account that triggered it?
- Is it jumping to other media?
- Is it continuing over a long period of time?
- Can we identify a connection to offline events relevant to the conflict?

Determining attribution

- Is it clear who posted first? Are others posting following the first poster?
- To identify paid trolls: Did the people posting follow the target before this? Did they post about this topic before? Are they from a different country to the target, and do they seem to post harassment elsewhere? Do they have limited identifying information and few friends? Have they had content removed from their profiles recently? Are they very prolific?
- To identify bots: Are a significant number of the accounts posting bots (e.g. use botomoter to find out)?

3. Evidence for contextual changes that signal affective polarization

The amplification of harassment and manipulation on social media results in **contextual changes that are important to understanding conflict dynamics**. These contextual changes are things that happen to people on social media. Contextual changes are a network effect that results from the interaction of platform design with human psychology – it cannot be attributed but it does need to be understood. Because contextual changes are not easily attributable, they require more proactive strategies to reverse pervasive behaviors or beliefs, such as narrative change campaigns or dialogue-based initiatives. Addressing contextual changes may be of more interest to people working on broader social cohesion or peacebuilding programming.

Contextual change	**Types of behaviors**	**How to find it**

Attitude polarization: perceptual shifts towards stereotypes, dehumanization, deindividuation and vilification of the “other”	Generalized traits attributed to a group, portraying them in a negative or derogatory manner, including comparing them to objects, animals, or inanimate things – and this content receives positive reactions Blame attributed to or fear expressed about a group, often through, often through the use of plural “you” and “they” pronouns to refer to a group – and this content receives positive reactions	Look for keywords or hashtags that refer to groups or use group signifiers. In this content, look for: - Negative sentiment (using a language sentiment model) - Use of words, phrases or graphics that attribute negative traits to a group (using a human or automated classifier) - Use of blame language in this content or in reply threads to it (using a human or automated classifier) Look for the most frequently used words after terms that refer to a group (word collocation) Look for the words used around terms that refer to a group (word embeddings)
Interaction polarization: reduction of quantity and deterioration of quality of meaningful communication across groups	Network segregation into group clusters with limited interaction When intergroup interaction happens, it is dominated by certain issues Less direct communication between groups, including blocking / unsubscribing from out-group content and unfollowing out-group users Movement to alternative, partisan social media platforms	Make a network graph of impressions, comments or replies (depending on what data is available) to look at network segregation Using the graph, find content that attracts comments / replies from two sides of a divide and identify patterns in discussed topics Less direct communication and movement to alternative platforms cannot be measured on platform – use some kind of poll or survey of platform users
Interest polarization: specific issues of contention give way	Sharing of information based on the source being in-group,	Look for keywords and hashtags that reference

to more general, simplified and unspecified claims	regardless of veracity (connected to spread of misinformation among in-group networks) Shaming of users who express or represent nuanced or complex positions Increased reach of high-emotion partisan calls-to-action	morality or moral authority in relation to the position of one's group Use a human or automated classifier to look for content which exhibits high-emotion calls to action or shaming of users who express nuanced views Where accounts can be classified as belonging to a particular group, look for the main narratives / topics in each group to identify similarities / differences
Affiliation polarization: aggregation of actors from formerly neutral, adjacent, or cross-cutting positions into a limited number of adversarial groups with increasing in-group cohesion	Patterns of exclusive content resharing from in-group members, including partisan media consumption and sharing Reliable validation for in-group defining content from a subset of accounts Reduced influence of moderate / bridge-building accounts v. increased influence of schismatic accounts Spiral of silence / self-censorship from non partisan accounts	Make a network graph of shares or mentions (depending on what data is available) to look at patterns of content resharing / validation (in some cases, filter this graph by topics of contention) Using the graph, find bridging accounts and accounts on the poles of the network, and consider their relative influence Self-censorship cannot be measured on platform – use some kind of poll or survey of platform users
Norm polarization: formation of new combative norms of interaction that displaces empathy or curiosity and reifies the erosion of trust between people	Acceptance of harsh tactics: hate speech, harassment, doxxing Expectation of negative outcomes when interacting with a group, including giving	To measure the acceptance of harsh tactics, look for comments that challenge a post that exhibits attitude polarization – when such comments are not present, norms may also be polarized

	up on possibility of constructive dialogue Expectation of the lack of veracity or value of content shared by the other group, including virality of outrageous claims about a group Expectation of incivility in intra-group interactions, including offensive discussion strategies, rapid position-taking and clapbacks	Shifts in expectations cannot be measured on platform – use some kind of poll or survey of platform users

Identifying whether the behavior is pervasive / having an impact

- For attitude, norm and interest polarization: measure the prevalence of content relevant to this type of polarization, and whether there is a growth in prevalence over time
- For interaction and affiliation: measure the degree of sorting on the relevant network graph, and whether there is more sorting over time
- In all cases: determining whether these changes online are impacting offline dynamics is very difficult, including with survey instruments – our best bet is to make a theoretical connection grounded in an understanding of how conflict escalation works (see this blogpost or this paper, for example).