Selection of data from online platforms that would enable better understanding of disinformation online and efforts to counter it

UN Interagency Dialogue on Disinformation and Data Transparency

Note: *This document is a draft for discussion purposes that has not been formally edited or reviewed. As such it does not necessarily represent the official views of the United Nations or the agencies of the staff who have contributed.*

March 2021

Through a series of informal dialogues aimed at enhancing UN cooperation in countering disinformation (understood as false or misleading content that can potentially harm human rights, and including misinformation and mal-information), staff from 21 entities from the UN family have collectively identified the following data points as important for research and actions in this area.[1]

Greater transparency from internet communications companies would allow the UN to more effectively understand and counter disinformation on COVID-19 and other key areas of public interest.

In regard to data access, the right to privacy should be respected but not serve as a means to unduly restrict such access. Where data is end-to-end encrypted, metadata about flow patterns may give insight. The UN is guided by the UN Principles on Personal Data Protection and Privacy and the United Nations Sustainable Development Group (UNSDG) Guidance Note on Big Data for Achievement of the 2030 Agenda: Data Privacy, Ethics and Protection.

For each of the data points below, disaggregated data on individuals and groups would be important, wherever possible, by: platform; device; date; location; language; gender; device; and behavior (frequency & recency, engagement, reach, volume & velocity), where relevant.

Other data points of disaggregation, where these exist, could include: age, country, state/province, city, neighborhood, vulnerable groups (migrants/refugees, indigenous peoples, ethnic & religious minorities, LBGTQI).

1. General data about content, mis and disinformation (and/or narratives of such content)

To understand the spread of disinformation online, platforms should make the following information available proactively or on request by UN actors and/or other parties with legitimate interest, such as researchers and journalists:[2]

[1] The informal interagency dialogues were convened by UNESCO and WHO, with the participation of EOSG, IOM, ITU, OHCHR, UN DESA, UN DGC, UNDP, UNEAD, UN ECA, UN ECLAC, UN EDA, UNEP, UN ESCWA, UN Global Pulse, UNICEF, UN Office of the Envoy on Technology, UN OSAPG, UN Women and UN75.

[2] Depending on data governance, regulations, legislation, data privacy, business secrets and other factors, the different datasets described above could be shared under through different means. Based on existing literature about data sharing modalities [see *https://www.nature.com/articles/sdata2018286*] , there are a number of options available for each of the indicators and metrics described above. Some of the options to consider are public data release, limited data release to certain organizations, release of pre-computed indicators or a question-and-answer model for specific cases.

- The number of posts and messages containing mis or disinformation detected in a selected time frame, and mechanisms which could allow an estimate of its representativeness - e.g. what percentage of the total content (by number of accounts, by daily traffic, etc.) that number represents.

- The number of users who saw, engaged with, reacted to, and shared/forwarded those posts and messages, and mechanisms to infer their representativeness.

- Detection of source of the content, particularly in regard to it originating on (or referencing/linking to) other platforms and/or services.

- Metadata needed to understand content 'mutated'/ adapted from past occurrences (zombies) or across content types and platforms.

- The number of users who saw those posts and messages, and what percentage of the total user base that number represents.

A categorization and quantification of the types of disinformation that are most commonly detected.

- A categorization and quantification of the types of mis and disinformation content formats (e.g., videos, photos, videos, text) which are most common and receive the most engagement.

- A categorization and quantification of the main characteristics of the actors involved in misinformation and disinformation (e.g., individuals, bots, private organizations, governments, impossible to identify).

- Demographics of those engaging with disinformation/ misinformation, for example engagement broken down by gender, age, previous engagement with disinformation/misinformation.

Additional areas where data is needed:

2. Content moderation and curation

To understand the ways in which content moderation and curation policies are conceived and implemented, and how these policies have an impact on countering disinformation while elevating access to information and freedom of expression, internet communications companies should make the following information available:

- The operations of algorithmic recommendations and trending signals in regard to amplifying or demoting mis and disinformation.

- Policies and policy changes as per categories of content, including how misinformation/ disinformation is defined and who monitors/checks the criteria and evaluates possible biases.

- The pipeline and procedures used to flag posts as misinformation/disinformation, including both human and algorithmic steps and the relative weighting between the two.

- An objective benchmark for the performance evaluation of this pipeline which includes an assessment of the specificity and sensitivity of the detection system, in particular accounting for those steps of the process which are automated (e.g. what is the percentage of messages which is constantly being missed in the initial automated triage?).

- Regular evaluations of how the content moderation pipeline is working to detect and acto on misinformation/disinformation.

- A characterization and quantification of the number of actors including bots, fake accounts, and "super spreader" accounts that have been shut down for propagating disinformation.

- Percentage of flagged content that gets reviewed and removed for violating policy, time to removal of flagged misinformation/disinformation, and number of impressions/engagements/shares before removal.

- The actions taken once a post is determined to be false (changes in promotion of post, addition of warning labels, etc.)

- The number of posts and of ads containing misinformation/disinformation which have been labelled, blocked, and/or removed; the percentage of volume that these figures represent.

- The number of posts and ads, subject to the measures described above, which were later overturned on appeal.

- The number of post removals requested by governments, the number of governments submitting removal requests, requests by non-state actors for content removal, and the percentage of moderation actions led by automated systems. The number of inauthentic "super spreader" accounts that have been shut down for propagating disinformation. number of restrictions by governments of platforms, including the location, length and rationale of each shutdown.

3. Partnerships and promoting reliable information

To understand if and how companies prioritize and promote reliable information, these entities should make the following information available:

- Categorization and quantification of the dissemination, timing, reach and engagement of posts promoting high-quality information - e.g. fact checked posts, content from content authorities.

- A list of partnerships developed with news organizations, journalists, fact-checkers, and other providers of high-quality content, which allow them to be made more visible on the platforms' timelines, as well as any potential support (including financial) provided to those organizations.

- The number of news organizations, health authorities, etc, that were made more visible on the platforms' timelines.

- The number of misleading and inaccurate claims flagged to fact-checkers for further verification.

4. Advertising

To provide increased transparency, the companies should make the following data available:

- A categorization and quantification of advertising related to misinformation/disinformation, including the content-related factors mentioned above (e.g. ad quantity, format, representativeness, engagement, and sources)

- Metrics on advertising, especially political advertising (e.g. ad spend, advertiser, micro-targeting of individuals and groups with protected characteristics)

- Types of automated (programmatic) advertising (such as timing of ads; bidding for slots; A/B testing of variants) as a percentage of total advertising

5. Users

To understand the types of users who spread and/or are impacted by disinformation, the companies should make the following information available:

- % failing Benford's law (suggesting suspicious, inauthentic behavior)

- creation dates

- actors whose accounts are suspended, for how long, and who are deregistered from the service

6. Economic metrics

To understand the financial aspects regarding countering or profiting (directly or indirectly) from disinformation, the following data should be available:

- Funding as a percentage of corporate spending to factcheckers / factchecking activities

- Principles of payment for fact-checking

- Revenue generated through user data (e.g. access to targeted advertising) and any "sale" of user data to third parties

- Percentage of spending on ensuring data security ◆

[3] See for example what is currently available: *https://www.facebook.com/ads/library/?active_status=all&ad_type=political_and_issue_ads&country=US*