



United Nations
Office on Genocide Prevention
and the Responsibility to Protect

A Comprehensive Methodology for Monitoring Social Media

**to Address and Counter
Online Hate Speech**

September 2024

Contents

Foreword	4
Introduction	7
PROJECT BACKGROUND AND OBJECTIVES	7
WHAT IS ONLINE HATE SPEECH MONITORING?	10
USE CASES	12
Current Challenges in Monitoring Hate Speech Online	13
DATA AVAILABILITY	14
REPRESENTATIVENESS	15
BIAS AND AUTHENTICITY	16
PRIVACY, SECURITY, TRANSPARENCY, AND HUMAN RIGHTS DUE DILIGENCE	17
Recommended Methodology	19
HIGH-LEVEL SUMMARY	19
HOW TO USE THIS METHODOLOGY	20
STEP 1: PLANNING	21
Defining the context	21
Specifying the intended usage	22
Determining the relevant forms of online hate speech to monitor	23
Identify partners	25
Human rights due diligence	26
STEP 2: DATA GATHERING	27
Identifying sources of online speech	27
Collecting relevant data	28
Selecting and compiling training datasets for automated classification	30

STEP 3: CONTENT CLASSIFICATION	31
Parsing message content and metadata	31
Training models of online hate speech	32
Classifying new messages	34
STEP 4: DEPLOYMENT	35
Setting up monitoring and alerts	35
Maintaining and refining	36
SAMPLE PROGRAMS	37
Example 1: The minimally-technical approach	38
Example 2: Leveraging social listening tools	39
Example 3: An automated early warning system	40
Case Study: Costa Rica	42
CONTEXT	42
PROGRAM DEVELOPMENT	42
RESULTS	43
KEY TAKEAWAYS	43
Future Innovations	44
shared infrastructure and processes	44
improved violence prediction capabilities	46
The benefits and challenges of widespread AI	47
Conclusion	48
Appendix A: The United Nations Strategy and Plan of Action on Hate Speech	49
Appendix B: Glossary of Key Technical Terms	50
Appendix C: Assessment of Existing Research and Tools	52
SUMMARY OF PREVIOUS RESEARCH FINDINGS	53
ANALYSIS OF HATE SPEECH IDENTIFICATION METHODS	54
AVAILABLE TOOLS AND DATA	55
Existing tools and services	55
Data sources	58

Foreword

The last few years have seen an alarming spread of hate speech around the globe. Hate speech fuels discrimination, undermines social cohesion, erodes shared values and, in some cases, constitutes incitement to violence and a potent driver of conflicts. In the most serious cases, hate speech may also act as a trigger of serious crimes, including genocide, war crimes and crimes against humanity, as we saw with the Holocaust, in Cambodia, the 1994 genocide in Rwanda against the Tutsi and the 1995 genocide in Srebrenica, in Bosnia Herzegovina.

Today, social media has become an echo chamber for hate speech and incitement to violence, accelerating their spread at an unprecedented level and, on occasion, leading to real world harm and violence. Unfortunately, in many cases the victims are already the most vulnerable in society. Technologies also continue to evolve very rapidly and uncontrollably. In this context, much more is required to effectively counter and address online hate speech.

On 18 June 2019, the UN Secretary-General launched the UN Strategy and Plan of Action on Hate Speech. He directed that my Office, the UN Office of the Special Adviser on the Prevention of Genocide, acts as the UN focal point for the implementation of this Strategy, coordinating system-wide efforts to counter and address hate speech. This responsibility entails providing support to UN field entities, civil society, and Member States to develop context specific and national action plans to tackle hate speech in line with international human rights and standards.

One of the thirteen commitments of the UN strategy and Plan of Action is to stay abreast of technological innovations and encourage more research on the relationship between the misuse of the Internet and social media and the factors that drive individuals towards violence. In line with this, my Office organized since 2020 a series of roundtables with tech and social media companies on how to tackle online hate speech. And from this engagement, in July 2023, published a policy document, [Countering and Addressing Online Hate Speech: A Guide for Policy Makers and Practitioners](#). This document provides recommendations to tech and social media companies on how to tackle hate speech on their platforms holistically, including by partnering with Member States, civil society and the United Nations.

Building upon this long-term engagement with tech and social media companies, "A Comprehensive Methodology for Monitoring Social Media to Address and Counter Online Hate Speech" aims to set up a systematic and common approach for monitoring hate speech in full respect of international human rights law and standards and based on the [UN Strategy and Plan of Action on Hate Speech](#). It is envisaged to help the work of various UN entities, and all relevant stakeholders engaged in tackling hate speech, including by providing examples on how to do this practically. It also provides an opportunity to reflect on what areas would benefit for further research and action to advance the fight against online hate speech globally.

It is my firm hope that this methodology will mark our first iteration towards strengthening our collective action on addressing and countering online hate speech, while upholding our fundamental rights, including freedom of opinion and expression, as well as non-discrimination and equality. This methodology will remain as a living document, helping to pave the way for further research on how combat hateful narratives online.

Finally, I would like to thank all those who supported and informed the development of this comprehensive online hate speech monitoring methodology. I am particularly grateful for the Permanent Mission of the United Kingdom to the United Nations for providing resources[1]. I would also like to extend my gratitude for Dr. Andrew Therriault, who successfully led in the development of this methodology as well as to all UN colleagues who are part of the UN Working Group on Hate Speech[2], and civil society representatives and experts who contributed their critical insights. I encourage all relevant stakeholders to use and widely disseminate this methodology and to join in our collective efforts to further the understanding and tackling of this phenomenon and, in doing so, build together more peaceful, inclusive and just societies.



Alice Wairimu Nderitu

Under-Secretary-General and Special Adviser on Prevention
of Genocide to the United Nations Secretary-General

[1] This project was undertaken with funding support from the United Kingdom's Foreign, Commonwealth and Development Office. All views reflected in this report remain the responsibility of the authors.

[2] UN Working Group on Hate Speech is comprised of 19 UN entities, including: UN Office of the Special Adviser on the Prevention of Genocide; UN Executive Office of the Secretary-General; UN Department of Political and Peacebuilding Affairs; UN Department of Peace Operations; UN Office of the High Commissioner for Human Rights; UN Department of Global Communications; UN Alliance of Civilizations; UN Development Programme; UN Educational, Scientific and Cultural Organization; UN Women; UN Office of Counter-Terrorism; UN Office of the Secretary-General's Envoy on Youth; UN International Children's Emergency Fund; UN Office of the Secretary-General's Envoy on Technology; UN Department of Economic and Social Affairs; UN Office of the Special Representative of the Secretary-General on Violence against Children; International Organization for Migration; UN High Commissioner for Refugees; UN Development Coordination Office.

Introduction

This report introduces a standardized methodology for monitoring online hate speech, to identify, assess, and mitigate risks, including when it constitutes risks of genocide, war crimes, and crimes against humanity. This methodology is based on an extensive review of existing methodologies used for this purpose across academia, technology companies, governments, the United Nations, and NGOs, and synthesizes those approaches into a standard set of practices that best fit the use cases relevant to the UN and its partners. These standards will enable better comparison of patterns of online hate speech observed across countries and by different organizations, as well as to enhance understanding of the context in which speech is delivered and received.

Our research builds upon the extensive prior work of other UN entities that tackled the inherent conceptual and legal challenges of identifying online hate speech in different contexts. The specific focus of our report is to address the technical challenges of implementing a hate speech monitoring program for social media. This methodology is intended to enable more effective use of social media monitoring technology and foster the development of shared data resources, to promote greater understanding of online hate speech trends (including trends that might pose risks of genocide and related crimes) in a consistent and timely manner, in line with international human rights standards and the [UN Strategy and Plan of Action on Hate Speech](#).

PROJECT BACKGROUND AND OBJECTIVES

In 2014, the UN Office of the Special Adviser on the Prevention of Genocide published a new [Framework of Analysis for Atrocity Crimes](#), which outlined a series of risk factors for future incidents of genocide, war crimes, and crimes against humanity and provided guidance on how to use these factors to assess overall risks of such violence.[3] This [Framework of Analysis for Atrocity Crimes](#) identified hate speech as an indicator of risks of genocide and related crimes, as well as potential triggering factor, the most serious form of which may constitute incitement to genocide. Five years later, Secretary-General Antonio Guterres unveiled the [UN Strategy and Plan of Action on Hate Speech](#). [4] This initiative provided strategic guidance for the organization to tackle hate speech, addressing its root causes and drivers and effectively responding to real-world impact, in line with international human rights standards.

[3] Framework of Analysis for Atrocity Crimes: A tool for prevention, available at https://www.un.org/en/genocideprevention/documents/aboutus/Doc.3_Framework%20of%20Analysis%20for%20Atrocity%20Crimes_EN.pdf

[4] <https://www.un.org/en/hate-speech/un-strategy-and-plan-of-action-on-hate-speech>

PROJECT BACKGROUND AND OBJECTIVES



Many organizations have sought to track its occurrence, particularly as an indicator of potential violence. But in this rapidly evolving field, there is no established standard for the optimal way to consistently define and track online hate speech, and the challenges involved in such a process are substantial. (An assessment of prior research and tools is provided as an appendix to this report.) This situation has led to a lack of consistency in the methodology and tooling for tracking online hate speech and limited the capabilities of these programs. And with social media data becoming harder to access in recent years (particularly due to restrictions on the Twitter/X API[5] and the sunsetting of Meta’s CrowdTangle tool[6]), technical barriers continue to be a major obstacle to effective use of social media monitoring by UN entities and partners around the world.

This project was launched to remedy that situation and establish a standard methodology for use by UN entities and partners to track online hate speech. Our methodology builds on existing research and programs and uses the results of earlier efforts to identify the approaches best suited to the UN’s specific use cases. By implementing a consistent set of best practices, readers will have the knowledge required to develop their own programs to identify and counter online hate speech. With a common methodological starting point, participating organizations will also be able to coordinate the development of shared tools and resources. This will lead to analyses which are more comparable across different contexts, enabling better research to understand how online hate speech is linked to real-world discrimination and violence, including genocide and related crimes. The goal of this work is to enable better predictive capabilities that facilitate early interventions, to prevent and mitigate these outcomes in the future.

[5] Coalition for Independent Technology Research (Executive Board), "Letter: Twitter’s New API Plans Will Devastate Public Interest Research", <https://independenttechresearch.org/letter-twitters-new-api-plans-will-devastate-public-interest-research/>

[6] "Meta Pulls Support for Tool Used to Keep Misinformation in Check" by Davey Alba for Bloomberg.com, posted June 23, 2022. <https://www.bloomberg.com/news/articles/2022-06-23/meta-pulls-support-for-tool-used-to-keep-misinformation-in-check>

Our report takes as its starting point the UN’s existing definitions of hate speech (as defined by the UN Strategy and Plan of Action on Hate Speech) and incitement to discrimination, hostility, or violence (as defined by the ICCPR[7], ICERD[8], and Genocide Convention[9]). Readers should pay close attention to the distinction of these terms, which has implications for monitoring, addressing and countering hate speech in line with international human rights norms and standards, in particular the right to freedom of opinion and expression; while hate speech is a broad category that includes many instances of legally protected speech, incitement has a clear legal definition and is prohibited under international law. For more information on these concepts, the UN has covered these topics in much greater detail in the UN Strategy and Plan of Action on Hate Speech and related documentation.[10] While the guidance herein seeks to apply to a wide range of applications and situations, we cannot practically address all possible scenarios in a single document, and instead have chosen to focus on specific use cases that we expect to be most widely applicable.



It is also worth noting that some programs combine online hate speech monitoring with efforts to track misinformation, disinformation, and other types of online content. Those programs generally leverage shared tools and processes to collect, analyze, and report on social media messages of various kinds and present a holistic picture of the information and opinion landscape. The methodology in this report is specific to online hate speech monitoring and does not address those other purposes but is designed to be readily compatible with such a multifaceted program as needed.

[7] <https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights>

[8] <https://www.ohchr.org/en/instruments-mechanisms/instruments/international-convention-elimination-all-forms-racial>

[9] https://www.un.org/en/genocideprevention/documents/atrocities-crimes/Doc.1_Convention%20on%20the%20Prevention%20and%20Punishment%20of%20the%20Crime%20of%20Genocide.pdf

[10] See in particular “Detailed Guidance on Implementation of the UN Strategy and Plan of Action for United Nations Field Presences”, available at <https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml>

WHAT IS ONLINE HATE SPEECH MONITORING?

Monitoring social media for hate speech is a process that combines three distinct tasks:



1. Collecting and parsing real-time social media data, to create a dataset of speech that is potentially hateful, along with relevant metadata such as the user who posted it and the amount of engagement (views, likes, reposts, etc.) it received.



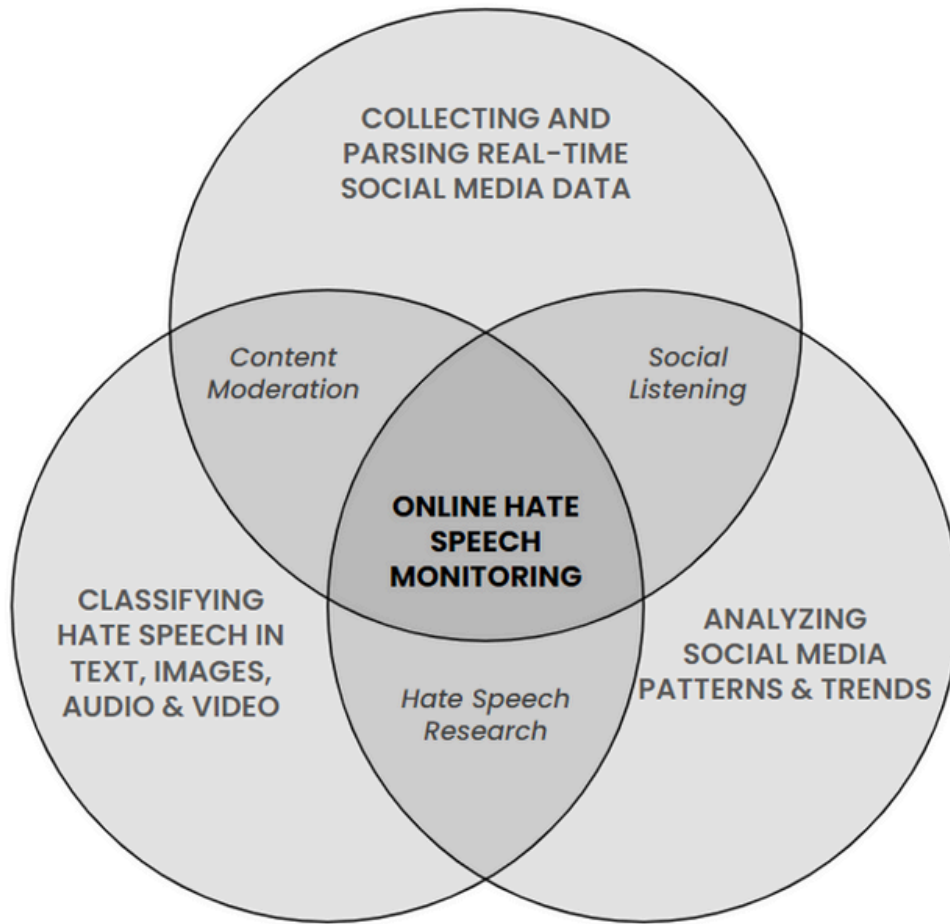
2. Classifying hate speech in text (and potentially also images, audio, and/or video) from this dataset to detect specific instances of online hate speech.



3. Analyzing social media patterns and trends in hate speech in terms of time, geography, and actors, to understand the potential offline implications of this speech.

This specific combination of tasks is unique to online hate speech monitoring, and both the existing methodologies and tools available are quite limited. However, other combinations of these individual tasks show up in other fields, and work in these fields can be leveraged to inform our methodology (*Figure 1*).

Figure 1: The Tasks Involved in Monitoring Hate Speech on Social Media



For example, social media platforms routinely classify user-generated content to flag speech that violates their platform policies, using a similar process to how we might categorize content as hate speech for our purposes. And likewise, commercial social listening tools developed for brand management are designed to track real-time social media content and report on patterns and trends, but these typically use simple filtering mechanisms such as keywords and hashtags rather than specifically classifying the type content (for example, as hate speech). Finally, many academic researchers have published studies on online hate speech patterns in a range of countries, but these typically take a retrospective view and are not designed to facilitate immediate response.

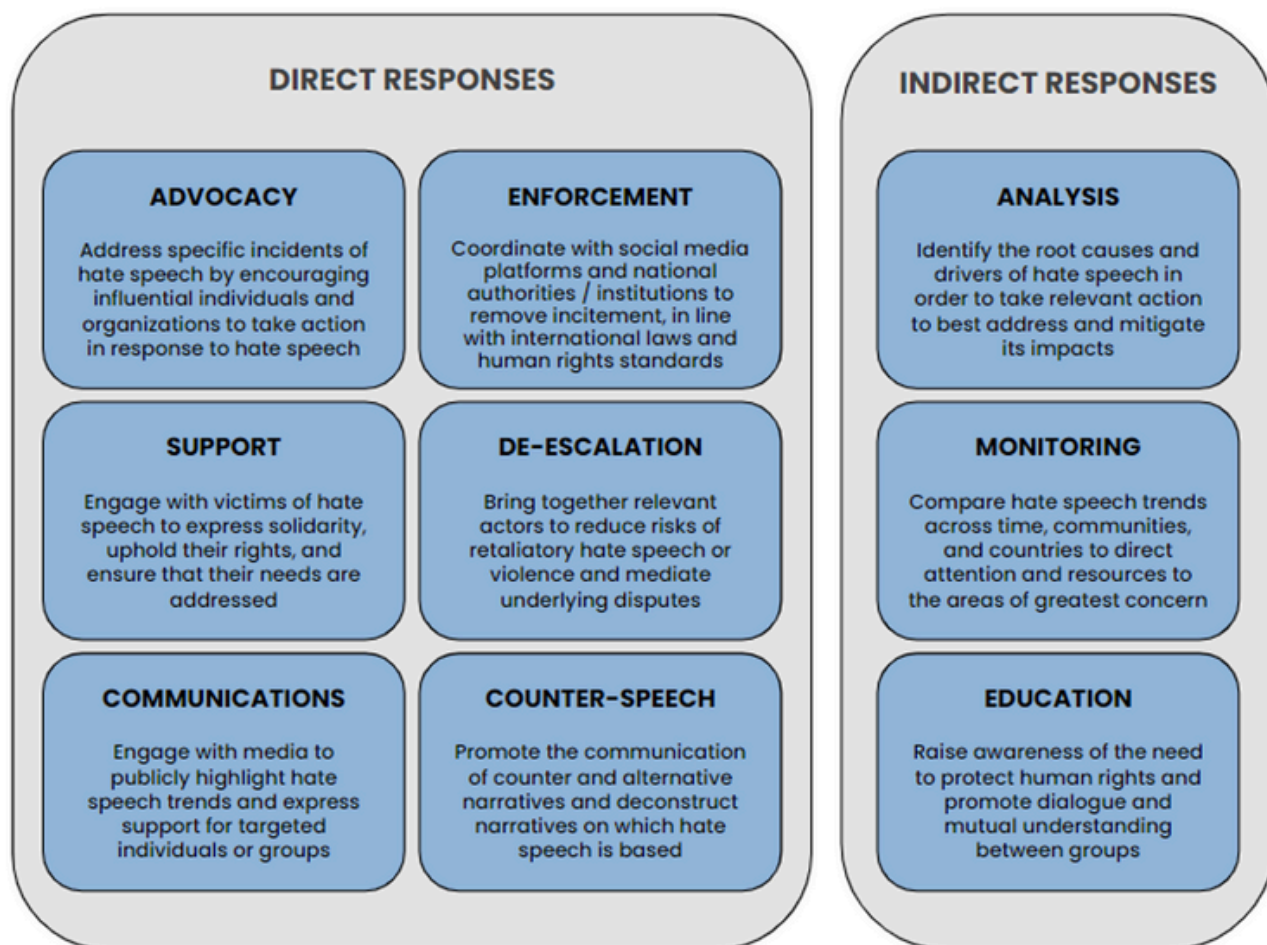
Because efforts in these other domains are more common than online hate speech monitoring, they have correspondingly received more investment in research and tool development. Throughout this document, we will leverage that work to support our own methodology and recommend strategies for applying those resources in the field. As of this writing, there are few off-the-shelf tools that fit all the needs for an online hate speech monitoring program, so in practice such a program will often require combining one or more existing tools with manual human inputs in order to accomplish all three tasks listed above. And where a single tool could meet most or all the listed requirements, the lessons learned from these other domains still help to inform the evaluation of possible options.

USE CASES

While there are myriad potential use cases for online hate speech monitoring, it is helpful to categorize these use cases based on the types of responses such a program is meant to inform. This is because different responses require looking at different sets of data and have different levels of required precision before acting, which in turn impose different requirements on hate speech monitoring programs. The [UN Strategy and Plan of Action on Hate Speech](#) identifies a wide range of actions that should be taken in response to hate speech, many of which can be informed by the results of online hate speech monitoring. Figure 2 provides a (non-exhaustive) selection of the most common potential responses that may be driven by such a program.

Figure 2: Examples of Direct and Indirect Responses to Online Hate Speech

(These examples are adapted from the 13 commitments and 27 actions listed in the [Detailed Guidance on Implementation of the UN Strategy and Plan of Action on Hate Speech for UN Field Presences](#), and do not include all possible responses to online hate speech.)



Direct responses involve addressing specific instances of online hate speech. This can include actions such as working with relevant local actors to promote effective responses, providing support to hate speech targets, and coordinating with tech companies to remove dangerous content which reaches the level of incitement (based on the six-part threshold test of the Rabat Plan of Action)[11] and must be restricted under Article 20 of the ICCPR.[12] These responses should be careful to account for the varying legal requirements imposed by different forms of hate speech: while instances of hate speech that reach the threshold of incitement are *legally prohibited* under international law, many other less-severe forms of hate speech still constitute *legally protected* speech.

Indirect responses, meanwhile, use the results of online hate speech monitoring to inform broader strategies and programs, enabling the UN and its partners to better fulfill their missions to prevent violence and protect human rights. Often these responses take place after hate speech has occurred, but the knowledge gained through observing these patterns over time can also help to identify the sort of events most likely to cause spikes in online hate speech, enabling more proactive interventions.

These potential responses were identified through a review of the [UN Strategy and Plan of Action's Detailed Guidance](#) as well as numerous discussions with individuals and organizations who have monitored online hate speech in the past, and this list is by no means exhaustive.[13] Online hate speech monitoring can provide value in any program or effort for which it would be helpful to have better visibility into patterns of online communications. In this regard, identifying online hate speech may be valuable both for enabling responses to hate speech itself and for using the prevalence and patterns of online hate speech as an indicator of the broader environment in a particular country or conflict.

Current Challenges in Monitoring Hate Speech Online

Online hate speech monitoring programs face many difficulties, some of which are intrinsic to these efforts and others which have only arisen in recent years. The subjective nature of hate speech makes consistent interpretation of hate speech a challenge - online or offline - and this is further complicated by the frequent usage of coded language and the potential for personal biases on the part of those doing the interpretation. But beyond these general challenges, there are issues specific to monitoring social media at the present time, and some of those issues are the focus of this section.

[11] The Rabat threshold test takes into account (1) the social and political context, (2) status of the speaker, (3) intent to incite the audience against a target group, (4) content and form of the speech, (5) extent of its dissemination and (6) likelihood of harm, including imminence (<https://www.ohchr.org/en/freedom-of-expression>).

[12] It is also important to note that excessive content removal could create chilling effects and undermine free speech, and it can disproportionately affect historically marginalized populations. Therefore, when monitoring hate speech we need to consider that researchers and civil society organizations play an important dual role in monitoring the presence and trends in hate speech in their communities as well as scrutinizing platform reporting and moderation policies.

[13] The OSAPG hosts a collection of related guidance and publications at <https://www.un.org/en/genocideprevention/publications-and-resources.shtml>, including "Plan of Action for Women in Communities and Countering" and "Addressing Online Hate Speech: A Guide for Policy Makers and Practitioners."

DATA AVAILABILITY

The most common challenge reported in nearly every discussion of online hate speech monitoring experiences is the difficulty of gaining access to social media data. This has always been a challenge to some extent, but it has grown more difficult as a product of recent developments, including:



The 2023 retirement of the classic Twitter API, which served as the backbone of social media monitoring and analysis for over a decade, in favor of a new X API that provided much more limited data at a substantially higher cost.



The impending retirement of Meta's CrowdTangle toolkit, which at one point enabled monitoring of posts across Facebook, Instagram, Twitter, and Reddit, but now only supports the first two platforms. First announced in mid-2022 but then postponed, it was finally announced in March 2024 that the platform will be shut down in August. (Another tool, the Meta Content Library, was simultaneously announced as an alternative, but how easily this new tool will fill the gap left by CrowdTangle's retirement remains to be seen.)



The shift in popularity from primarily text-based content on Facebook and Twitter to image- and video-based content on Instagram, YouTube, and TikTok. Not only are these forms of content more difficult to collect and store, but they are also much more difficult to automatically parse and analyze than text.



The rise in private group messaging on apps like WhatsApp, Telegram, and Signal, which make an increasing amount of online hate speech largely inaccessible for monitoring.

Across all our conversations with individuals and groups involved in online hate speech monitoring, both within the UN system and from NGOs and academic institutions, there was a universal call for increased data sharing and transparency on the part of social media platforms.[14] The ability of the UN and like-minded organizations to use online hate speech monitoring to prevent violence and discrimination depends on the cooperation of these technology companies. Without their participation, the harms caused by online hate speech on their platforms simply cannot be effectively mitigated.

[14] See also the UN Global Principles for Information Integrity, available at <https://www.un.org/en/information-integrity/global-principles>.

REPRESENTATIVENESS

The data we can collect, moreover, is only a limited snapshot of speech conveyed online. It is available only from a subset of platforms, and only represents a portion of the speech on each platform. We do not observe speech in private posts, messages, or groups. Nor do we see speech that is missed by keyword or hashtag filters, or that which does not come from sufficiently prominent accounts or get enough engagement to meet the relevance criteria set by some platforms for inclusion.

The platforms' own content moderation efforts also create a challenge for monitoring, as these programs evolve over time and seldom share information about how much content they reject or remove. This makes it difficult to compare patterns over time and across countries, languages, and platforms. That also drives the authors of online hate speech to adapt their language to avoid moderation, which similarly interferes with our ability to monitor. And because we do not often see the content that gets removed, restricted, or blocked before posting, we cannot observe a complete picture of the online conversation.

The representativeness of available data is especially limited when it comes to languages other than English and, to a lesser extent, a small number of other Western European languages. Moderation policies are often much less consistently enforced in these languages, particularly for languages where social media companies have not invested substantially in hiring native speakers as moderators. The same issue is also seen across the existing training data, dictionaries of hate speech vocabulary, linguistics, and tools for online hate speech detection.

Additionally, the vast majority of the research on and monitoring of hate speech on social media platforms has focused on the USA and Europe. This leads to a gap not only in tools and data, but in the understanding of the extent and dynamics of the spread of hate speech in other regions. This gap remains crucial to bridge given the inherent contextual nature of hate speech.[15]

[15] Perini, Reja et al, "Monitoring and Addressing Hate Speech on Social Media - contemporary challenges", UNESCO, Paris, June 2024.

BIAS AND AUTHENTICITY

Recent studies on online hate speech identification have also noted that tools and processes for categorizing hate speech can produce biased results when classifying messages from or about certain groups. This is especially problematic when relying on specific keywords to screen for online hate speech, since the same words can have very different interpretations depending upon the speaker and context. Such bias can lead to both false positives (when non-hate speech gets flagged by moderation or monitoring systems) and false negatives (when actual hate speech goes unaddressed and unnoticed) and muddle our overall picture of the online hate speech environment.

This problem is relevant for the detection of online hate speech of all different types, and particularly when focusing on hate speech that reaches the level of incitement, which is prohibited by international law. These cases emphasize the indispensability of diverse and well-trained human reviewers before acting on instances of detected hate speech. At the same time, we must acknowledge that human reviewers are also fallible and subject to cognitive biases of their own. This situation must be accounted for when performing human rights due diligence assessments of any monitoring program, and appropriate safeguards and training should be incorporated into these programs to minimize any adverse impacts.

Data on online hate speech patterns can also be skewed by the presence of bots and fake accounts. In some cases, these accounts (which often operate in concert with each other) can artificially inflate the engagement and reach metrics of hate speech content, making these messages appear more noteworthy than they are. At the same time, these networks can have a real-world impact, when their utilization of the platforms' algorithms successfully amplifies the visibility of hate speech messages beyond what they would receive from organic engagement alone. Likewise, fake accounts can be used to impersonate prominent figures and groups to lend online hate speech false credibility, while at the same time such tactics can also be used to hide the identities of online hate speech authors and provide them with a veneer of deniability. Altogether, the impacts of inauthentic accounts serve to further cloud our view.

PRIVACY, SECURITY, TRANSPARENCY, AND HUMAN RIGHTS DUE DILIGENCE

Monitoring social media posts, even those posted on public websites, creates a responsibility on the part of the organization doing the monitoring to ensure that the information it collects is managed and used appropriately and does not cause harm.

To protect the privacy and other rights of individuals reflected in that data (which includes not only the authors, but also others mentioned or otherwise tied to the conversation), online hate speech monitoring programs should follow the guidelines adopted by the UN HLCM in the Personal Data Protection and Privacy Principles in 2018.^[16] These include only utilizing personal data for a specific permitted purpose, only retaining data for as long as required to serve that purpose, and limiting data collection and usage to only that which is necessary for that purpose. Such programs should also recognize that deidentification measures (such as not recording social media users' account names) are often inadequate to fully anonymize data, so any monitoring program must treat all social media data as personal and treat it accordingly.

Collecting such data also creates potential security risks. Even data that was posted on a public website should be protected as private data would be, because that information may no longer be public (for example, if the post was later deleted or if the user changed their privacy settings). Its collection and analysis may also exacerbate the risk of misuse or retaliation, because being compiled by the UN and labeled as potential hate speech can change the interpretation of a given message by other parties. And likewise, programs must take into account potential safety and security risks this may cause for UN entities, their staff, and other partners. These risks include but are not limited to reputational risks, cybersecurity risks, personal safety risks, and risks to the organization's ability to fulfill its mandate.

[16] https://archives.un.org/sites/archives.un.org/files/_un-principles-on-personal-data-protection-privacy-hlcm-2018.pdf

Online hate speech monitoring programs also need to consider how best to maximize transparency while protecting personal data and the integrity of the program. These programs should generally be as transparent as possible about their systems and methods, while also keeping in mind that excessive disclosure could enable manipulation by those who want to avoid attention, much like users who often circumvent moderation efforts on platforms. Likewise, systems which employ automated classification models inevitably make some mistakes (which is why human review is necessary before direct responses are taken) and sharing these raw outputs may not be advisable if they could be misrepresented by those seeking to undermine confidence in the program. Limitations on legal collection of personal data also represent a challenge in many countries, and steps should be taken to ensure compliance with relevant laws. Though many regional, national, and local data privacy laws contain exceptions for scientific research and protection of public safety, these must be evaluated on a case-by-case basis. Local laws may also present challenges in countries where speech is restricted to a degree greater than under international law, particularly when some monitored speech might be considered to be a threat to safety or security.

Further, the conduct of online hate speech monitoring should be informed by and compliant with international human rights norms and standards, and particular attention be paid to avoid any undue restrictions of freedom of expression.^[17] Human rights due diligence should be conducted to identify the risks these programs may pose to people (including those conducting the monitoring) and all reasonable steps taken to prevent or mitigate such risks, in accordance with existing guidance. Effective due diligence would need to be a continuous, ongoing and iterative process; supported by efforts to embed human rights into policies and management systems; and aimed at enabling programs to remediate adverse impacts potentially caused.

[17] Refer to UNESCO Guidelines for the Governance of Digital Platforms (<https://unesdoc.unesco.org/ark:/48223/pf0000387339>) which aim to safeguard freedom of expression and access to information while dealing with content that could be permissibly restricted under Article 20 of the ICCPR; In this context, see also the Rabat Plan of Action.

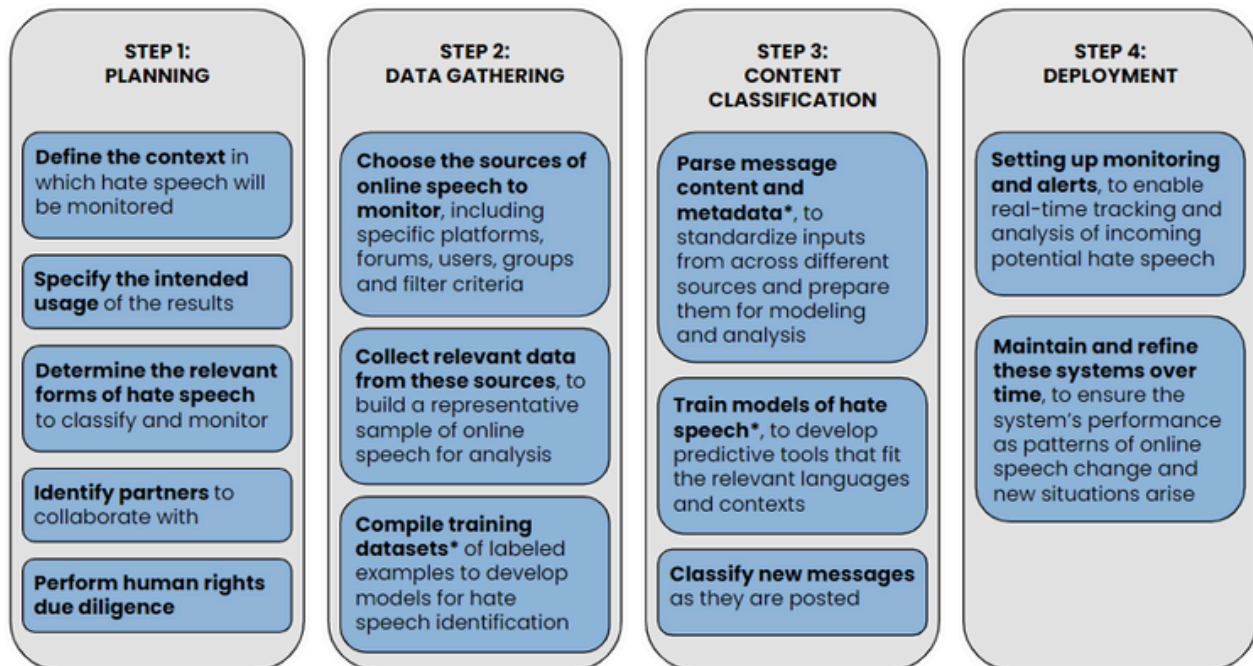
Recommended Methodology

This section distills the assessment of existing research and tools into a set of recommended practices for monitoring online hate speech. The exact implementation of these practices will vary across contexts because the choice of the most appropriate tools and workflows depends upon factors such as language, intended usage, technical resources available, and patterns of social media usage in a given area. But with these best practices in hand, users will be able to implement new applications in a consistent way that leverages the most up-to-date technology and approaches. Long-term, these practices are also designed to facilitate greater collaboration and development of shared tools and resources across organizations.

HIGH-LEVEL SUMMARY

Monitoring online hate speech is a process comprised of a series of discrete actions that are required to accomplish the tasks introduced earlier in this report (collecting and parsing real-time data, classifying content, and analyzing patterns and trends). We have broken down this general process into the required steps shown in Figure 3.

Figure 3: Required Steps for Monitoring Hate Speech Online



** Actions marked with an asterisk are mainly needed for automatically processing and classifying hate speech and may be skipped if not applicable.*

A wide assortment of online hate speech monitoring programs can be developed that fit this methodology, from a low-tech process of manual monitoring to the latest AI-driven big data platform. The main value of this framework is that it provides a consistent and straightforward way to implement these programs, regardless of the specific program details or level of automation. Each step in this process is essential for UN entities and partners to effectively address online hate speech and use it as an early-warning signal for potential discrimination and violence.

HOW TO USE THIS METHODOLOGY

This methodology is designed to serve as a guide for individuals and groups tasked with or considering an online hate speech monitoring program, regardless of their level of technology expertise or the resources available. It not only lays out a roadmap which details the key steps in developing such a program, but also provides specific recommendations on how to best execute those steps in a manner that aligns with both UN policy and scientific best practices. It was written to be especially beneficial to program staff with minimal technical background, but it will likewise be informative to technical experts who work with these systems, both within the UN system and without.

The individual steps shown in Figure 3 each contain a number of different elements, and many of them may be bundled together as part of off-the-shelf tools built by external researchers or vendors. For example, if a commercially available tool has a pre-trained hate speech classification model built in, that may help in fulfilling the requirements Step 3 (though manual verification of some or all flagged content may still be required). But even when much of the technical work leverages existing systems or tools, this methodology can inform how those systems and tools are selected, implemented, and interpreted in the field.

The sections that follow break down these steps into specific practices that together give a comprehensive methodology for monitoring online hate speech. Each section provides guidance for designing such a system and choosing its key parameters, but every individual implementation will require a unique approach tailored to meet specific program requirements. Collectively, a system that successfully implements all of these practices will have all of the elements required to achieve its objectives.

STEP 1: PLANNING

Before any online hate speech monitoring effort begins, those designing the program need to answer a series of critical questions that inform the technical choices to be made during implementation.

Defining the context

The first challenge is to define the context in which online hate speech is to be monitored. Generally, a monitoring program will be developed to fit a particular place or situation, because the specific characteristics of hate speech and available responses to it vary across contexts. Key questions to answer include:

- ▶ What geography is of interest?
- ▶ What language(s) and linguistic contexts does it need to support?
- ▶ Are there specific topics it should focus on?
- ▶ Are there specific targets of likely hate speech it should pay attention to?
- ▶ Who are likely to be the most influential producers or promoters of hate speech?
- ▶ Where on the internet is hate speech most prevalent and dangerous?




In many cases, the answers to these questions may seem obvious in light of observed cases of hate speech or ongoing conflicts. But programs should also consider possible answers that are less immediately clear. For example, many cases of online hate speech which occur in the context of religious or ethnic strife also target women, migrants, refugees, asylum-seekers, internally displaced persons, and/or the LGBTIQ+ communities. As such, the prevalence of online hate speech directed at these groups should also be considered when defining the scope of the monitoring program.

When defining the context of a monitoring program, it can also be helpful to gather verified examples of online hate speech in that specific context, to begin building a shared understanding of the situation and better answer the questions above. Such examples may be readily available from UN country staff, external partners including civil society and affected communities, media sources, or other organizations. Programs need to be careful not to define the context too narrowly based on a limited set of early examples, thereby inadvertently excluding other types of relevant hate speech, but used appropriately these examples can be illustrative of the unique characteristics of online hate speech in that particular setting.

Specifying the intended usage

With the context in mind, the next factor to consider is the range of possible actions that could be taken once online hate speech is observed. A general categorization of potential responses was introduced in Figure 2, but for a more detailed list of potential responses, refer to Section IV of the [Detailed Guidance on Implementation of the UN Strategy and Plan of Action on Hate Speech for United Nations Field Presences](#). While it may not be possible to identify every use case for an online hate speech monitoring program in advance, the initial plans should note all likely use cases based on the program's current situation and mandate. Those discussions should also include other UN entities and partners who are working on related efforts and who may wish to leverage this data in the future.

In discussions with UN staff and partners who have engaged in online hate speech monitoring, three recommendations were especially noteworthy:

-  First, the intended usage should be chosen to enable the program to make a tangible impact in the near term. This will help validate the program's effectiveness and build momentum for long-term success.
-  Second, monitoring programs should also use this opportunity to set clear expectations with relevant stakeholders (both internal and external) about what the program is and is not able to accomplish.
-  Finally, these programs should outline what will happen in the event that particularly extreme cases of online hate speech (such as potential incitement to discrimination, hostility or violence or genocide) are detected. Even if the program is not intended to address these situations directly, there should be a protocol in place for how such incidents can be escalated through the right channels to ensure that there is appropriate follow-up. If such a process is not feasible in a given context, that limitation should be clearly conveyed to all stakeholders to avoid creating incorrect expectations.

Determining the relevant forms of online hate speech to monitor

In order to accurately classify online hate speech, every program needs to have a clear working definition of hate speech against which to evaluate content. The [UN Strategy and Plan of Action on Hate Speech](#) defines it as follows:

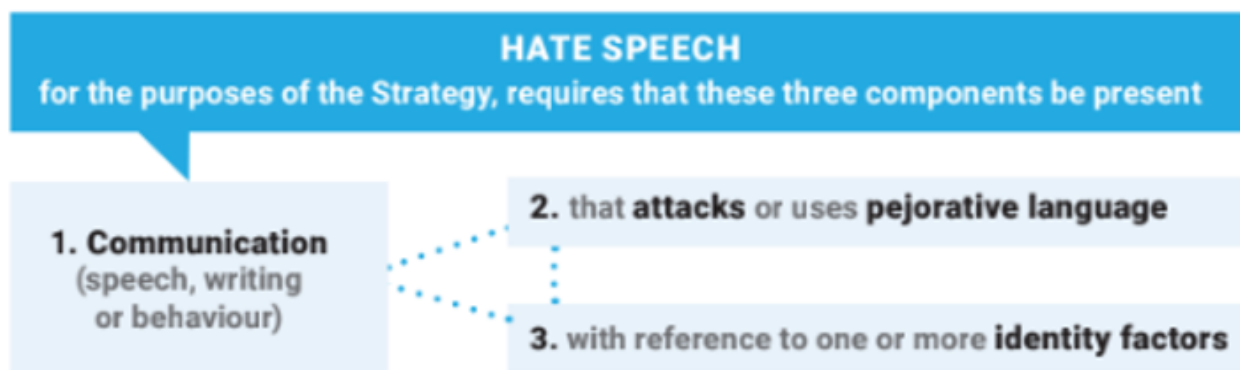


Hate speech: “any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor”

This definition includes three main components: it is (1) communication (2) that attacks or uses pejorative language (3) with reference to one or more identity factors (see Figure 4). More detailed information about how to use this definition is available in the [Detailed Guidance on Implementation of the UN Strategy and Plan of Action on Hate Speech for United Nations Field Presences](#).

Figure 4: The Components of Hate Speech Under the UN Strategy and Plan of Action

(Reproduced from Figure 1 of the [Detailed Guidance on Implementation of the UN Strategy and Plan of Action on Hate Speech for United Nations Field Presences](#).)



Depending upon the context and aims of an online hate speech monitoring program, it may be sufficient to simply track all online hate speech that meets this general definition. In some circumstances, however, a finer-grained breakdown of hate speech into different forms is called for. The [Detailed Guidance](#) describes a three-tier classification system, which ranges from the least severe forms of hate speech (which are still legally protected speech) to cases of incitement (which are prohibited by international law).

This last form is particularly relevant to online hate speech monitoring programs, and is defined by the International Convention on Civil and Political Rights as follows:



Incitement to discrimination, hostility, or violence: “any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence”[18]

Specific criteria for incitement are given in the Rabat Plan of Action, which outlines a six-part test for whether a given statement constitutes incitement.[19] These criteria are important because monitoring programs which seek to use enforcement (either through platform moderation or legal processes) to directly remove online hate speech and/or sanction its authors must apply this test before acting against any instance of detected online hate speech to confirm its illegality.

The most severe category of incitement is incitement to genocide, which the 1948 Convention on the Prevention and Punishment of the Crime of Genocide (often referred to simply as the “Genocide Convention”)[20] defines as follows:



Incitement to genocide: “direct and public incitement to genocide”, where genocide is defined as “ any of the following acts committed with intent to destroy, in whole or in part, a national, ethnical, racial or religious group, as such: (a) Killing members of the group; (b) Causing serious bodily or mental harm to members of the group; (c) Deliberately inflicting on the group conditions of life calculated to bring about its physical destruction in whole or in part; (d) Imposing measures intended to prevent births within the group; (e) Forcibly transferring children of the group to another group.”

[18] Article 20 (2) of the ICCPR. See also article 4 International Convention on the Elimination of All Forms of Racial Discrimination, according to which States Parties “shall declare an offence punishable by law dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another colour or ethnic origin”.

[19] <https://www.ohchr.org/en/documents/tools-and-resources/one-pager-incitement-hatred-rabat-threshold-test>; The threshold test of the Rabat Plan of Action has been used by Meta’s Oversight Board in more than a dozen decisions (see [A/HRC/55/74](https://www.oversightboard.com/decision/), para. 63 and <https://www.oversightboard.com/decision/>).

[20] https://www.un.org/en/genocideprevention/documents/atrocities-crimes/Doc.1_Convention%20on%20the%20Prevention%20and%20Punishment%20of%20the%20Crime%20of%20Genocide.pdf

Incitement to genocide is an especially extreme case of hate speech, in that it requires the specific intent to physically destroy a national, ethnic, racial, or religious group, not just individual members of the group.[21] In some contexts, it may be important for a monitoring program to differentiate between incitement to genocide and other forms of incitement when categorizing instances of online hate speech, because the subsequent actions taken in response may be different.

Note, also, that most existing computer models of “hateful” or “derogatory” speech do not apply as strict a definition of hate speech as that given above, and thus may be of limited utility if the speech they include or exclude does not align with the program’s needs. This same caveat also applies to datasets used in hate speech research from academia, industry, or other organizations, so caution should be used if those datasets are employed as training data for automated systems of online hate speech detection.

Identify partners



Many online hate speech monitoring programs are conducted in multi-stakeholder partnerships which may involve civil society organizations (both formal and informal), affected communities, academia, national human rights institutions, governments, politicians, media actors, women’s groups, and others. These external partners can help to provide resources, offer unique skills, and take on portions of projects that are not a good fit for UN entities. Such partnerships can also help to directly build local support for countering online hate speech and engaging and empowering affected communities.

In some programs that involve enforcement actions against content that constitutes incitement, proactively establishing relationships with national government or other authorities and social media platforms generally results in more effective enforcement by the platforms than reliance on public reporting channels.

[21] <https://www.un.org/en/genocideprevention/genocide.shtml>

Human rights due diligence

The final stage of planning an online hate speech monitoring program is to evaluate its potential impacts on human rights, including privacy, security, and freedom of expression. This should be done in accordance with any Human Rights Due Diligence policies that are applicable to the UN entity engaged in monitoring, and even in the absence of such a policy, should consider the four key components of human rights due diligence:

- 1. Identifying and assessing the actual or potential adverse impacts.
- 2. Acting on these findings to prevent and mitigate these impacts.
- 3. Tracking the progress of these actions.
- 4. Communicating with key stakeholders.

This process should be an ongoing one that continues through the lifecycle of the program. While some of the potential impacts can be identified and addressed during the planning phase, others may not show themselves until later. As such, all programs should have plans in place to recognize and respond to adverse impacts on human rights, and the outcomes of these plans (for example, steps taken to remedy observed impacts) should be included in reporting about the program's outcomes.

Particular care must be taken to anticipate any potential unintended impacts of the program on those involved or the broader community. For example, several existing programs noted that repeated exposure to online hate speech by those doing the monitoring caused significant mental and emotional harms to those individuals. As such, programs that anticipate a potential risk of such harms should have plans in place to provide support to individuals affected and limit exposure to the minimum necessary to achieve the program's goals.[22]

[22] One example of how this could be done: rather than requiring human review of all flagged content, tech companies often use a technique called "hashing" to label potentially-duplicated content (for example, copies of the same image or video). When a given piece of content is identified as in violation of platform policies, that content's "hash" can be added to a banned list, so future postings of the same content can be automatically removed without the need to expose a human reviewer to each individual copy.

STEP 2: DATA GATHERING

With plans in place, the next step is to source the required data for analysis.

Identifying sources of online speech

To start, we must identify where our data on online speech is to come from, both for initial implementation and ongoing deployment. This involves the following tasks:



1. Choose which platforms and forums to monitor.

This should primarily be based on local usage patterns, and (if available) the known prevalence of hate speech and related content on each platform. Because not all social media platforms provide easy access to their data, this choice should also factor in the realities of data access and the resources available. Depending upon the context, it may also be worth monitoring forums that are not typically thought of as social media - for example, comment sections on news or other media websites. These can serve a similar function to social media platforms, especially around high-profile events, and in many cases have higher frequencies of hate speech than individual posts on public forums.



2. Choose specific users and groups to monitor.

In many cases, programs may wish to focus particular attention on specific users or groups (for example, political parties or leaders) to make sure they stay on top of the most relevant sources of messaging. These users or groups could include both the potential authors and potential targets of hate speech. Many of these selected accounts will be public figures or organizations, but it may also be worth including highly-influential private individuals or organizations if at least some of their content is relevant to the chosen context.[23] In other cases, though, it can be valuable to look at a broader set of users, since some programs have found that the majority of online hate speech is authored not by extremist leaders or activists but by private individuals. In some instances, past programs have even found it helpful to deliberately focus on inauthentic accounts, as these accounts may be used to produce hate speech content that is subsequently amplified by other users and groups.[24]

[23] These may be available from existing analysis done by the UN at the country level, but a new study may be required if this information is not available or does not include the most up-to-date information for the particular context of the program.

[24] In these instances, inauthentic accounts may be used as the originators of hate speech content in order to shield those promoting that content from the potential repercussions (such as platform enforcement or legal liability) they might face if they had originally posted it from their own accounts.



3. Choose filter criteria to use.

Because of the volume of social media content available, even automated systems need to set some kind of filtering mechanism when deciding which content to include. At minimum, some level of engagement threshold is almost always used, since the overwhelming majority of social media posts get little or no attention.[25] Beyond that, many programs - especially those which do not use automatic classification but monitor more than just a handful of accounts - will also want to filter content by keywords and/or hashtags, in order to narrow down the included content to that most likely to include relevant hate speech. The selection of these criteria is highly context-dependent, and finding a balance between including too much non-hate content and not catching all instances of online hate speech may need some adjustment before the right balance is found.

More information on identifying social media usage by country and region, determining influential users and groups, and choosing the most relevant keywords and hashtags can be found in the DPPA's Social Media Analysis guide (which is not specific to hate speech, but is intended to be relevant to that usage).[26]

Collecting relevant data

Once these data sources are chosen, the first technical challenge is to gather social media content for analysis. There are many options for how this could be done, but they fall into four basic categories:



Manual data collection.

The most basic approach to collecting social media data is to have individual staff or partners gather it by hand - in other words, to simply use the social media platforms and forums being monitored. This typically involves setting up dedicated accounts on each platform, then following selected users and groups, and/or using built-in search functions to look at specific keywords, hashtags, pages, or other locations where online hate speech may be found. When likely hate speech is found that meets the chosen criteria (for example, based on its author or level of engagement), it is then recorded in a centralized, secure location for later analysis and potential response. These records should at minimum include the exact text, author, time, and date of the content, as well as a link to the source, but ideally would also include a screenshot or other reproduction in case the content becomes no longer available in the future.

[25] For example, a 2018 report by Mention found that most Twitter posts receive exactly 0 likes and 0 retweets. Since engagement is closely correlated with visibility, such posts are rarely seen and probably not worth responding to or including in analyses of the overall social media conversation.

[26] <https://reliefweb.int/report/world/social-media-analysis>



Automated data collection.

A broader data gathering effort will automatically collect relevant content in a database for analysis. This typically involves setting up an automated process that queries an application programming interface (API) made available by the platform itself or a third party, either publicly or privately. In some cases, this could also involve creating computer programs that record content displayed on the normal user interface (“scraping”), but this is generally not recommended because most platforms prohibit such usage in their terms of service.



Third-party applications.

Because manual data collection is highly labor-intensive and automated data collection carries substantial technical costs, a common compromise solution is to use third-party applications to collect and triage social media data. These applications - which may be available from commercial vendors, NGOs, or the platforms themselves - leverage shared infrastructure to make social media data accessible without the need to develop a custom tool. This capability is often bundled with other features that allow users to search, analyze, track, and even classify new content automatically. The tradeoff for this convenience is that these applications typically include their own content filtering rules[27] and restrict access to the underlying data, which limits their utility for custom analysis. And because most such applications are not specifically built for tracking online hate speech, they are not typically sufficient for a broad-based tracking program that relies on automated classification.[28]



Crowdsourcing.

Many online hate speech monitoring programs will want to utilize the efforts of UN staff, NGO partners, and others to flag specific instances of potential online hate speech. This can be useful for both initially gaining an understanding of the nature and patterns of hate speech and also for ongoing monitoring. If this channel is made available to members of the public, it is often helpful to have this information relayed through an external partner rather than to the UN directly, with an understanding about how this information will be filtered and triaged.[29] It is also important in these cases to set clear expectations for the public about what follow-up can be expected after a report, as the receiving organization may not have the capacity or authority to react to reports in the way the public might hope for.

[27] For example, Meta’s CrowdTangle platform only collects data from accounts and pages that meet their own criteria for popularity and engagement.

[28] Though some third-party tools do enable automated tagging of hateful content, to our knowledge none use models trained on the UN’s specific definitions of hate speech or incitement, so they may still be inappropriate for this task. At most, such models should only be used for approximate estimation of hate speech levels, and not for any analysis requiring precision or for responses that do not first involve a manual verification of all flagged speech.

[29] This is especially important in situations where such reports might become politicized and subject to deliberate over-reporting in an attempt to suppress opposing speech or interfere with monitoring. This is similar to the problem social media companies often face when their trust and safety tools are subject to abuse. In these cases, the potential resource burden of reviewing and responding to all reports can become extreme. This could in turn cause legal complications if the receiving organization (such as the UN) may be obligated by law to address each individual report.

For programs tracking speech on multiple platforms, a combination of these approaches may be required. In such cases, manual processes may be required to supplement more readily available data from other sources.

Selecting and compiling training datasets for automated classification

Finally, programs that will utilize automatic classification of online hate speech will first need to compile training datasets of labeled examples. As guidelines for developing training datasets:



These should include samples of real messages taken recently from the same context as the speech to be tracked, which have been labeled as either hate speech or not by human reviewers using the definition selected previously.



For programs that seek to track online hate speech across multiple languages, separate datasets should be created for each language, as patterns of hate speech are often unique in each language and can be missed when using automated translation tools.



Training datasets must include a mix of both hate speech and non-hate speech and will typically include several thousand labeled examples or more. Smaller datasets may be acceptable if used in conjunction with pre-trained models and/or existing datasets not specific to this context, but larger datasets will enable much more accurate classification.



If desired, messages determined to be hate speech may further be classified into different levels of severity or types of hate speech (for example, targeting a specific ethnicity or gender), as such labeling may help to improve the underlying models (which will be trained in the next step).

Compiling an original training dataset is recommended for every program that uses an automated classification model, even when existing datasets are available from academic researchers and other sources, because those datasets will almost always have used a different definition of hate speech and be based on messages that are at least several years out of date. If these datasets are reused, at minimum, their labels should be manually reviewed for consistency with the UN's definition of hate speech (and specific levels). Furthermore, they should be supplemented with a substantial amount of new content (both hateful and not) so that changes in normal language patterns over time are included in the training data.

In addition to labeled training examples, automated classification models also benefit greatly from having access to lexicons of specific terms commonly used in hate speech content. There are several publicly available sources for such data[30], in a wide range of languages, but it may also be worthwhile to develop an original dataset containing the most up-to-date and context-specific terminology seen in recent examples.

STEP 3: CONTENT CLASSIFICATION

Once social media data has been gathered, the next step is to classify whether or not each piece of content is hate speech.

Parsing message content and metadata

Each message needs to be parsed in order to prepare it for classification. For automated systems, this step requires extraction of key information from both the content itself and other descriptive data about the message (its metadata). The specific set of data to parse will vary based on the context, the platform and type of the message, and the modeling approach that will be used, but common features may include:

- ▶ Presence of relevant keywords and hashtags.
- ▶ Inclusion of images, audio, or video.
- ▶ Mentions of other users or groups.
- ▶ Other messages being reposted or replied to.
- ▶ Links to other websites.
- ▶ Date and time.
- ▶ Geolocation.
- ▶ Language.
- ▶ Engagement (likes, retweets, replies, etc.).
- ▶ Account characteristics (age, followers, posts, etc.).

[30] See <https://hatespeechdata.com/#Keywords-header> for a list.

More advanced hate speech classification models may even go a step further and include information about the author's social network (sometimes called graph data), as these details have been shown to improve models' ability to accurately categorize online hate speech, but such approaches should be taken cautiously. While in some cases this additional data can reduce bias in the resulting model (by adding better context for each message)[31], it also greatly increases the model complexity and may reinforce biases or errors in the training data.

For programs which classify online hate speech manually, this level of message parsing is not necessary, but we still recommend collecting some additional data points if available. For example, the Rabat Plan of Action's six-part threshold test for incitement includes as one of its criteria the "extent of the speech act", including its reach and the size of the audience. In the case of a social media message, the level of engagement is key to determining the speech's severity, so that information should be available to the person(s) tasked with determining whether a given message constitutes incitement or another form of hate speech.

Training models of online hate speech

Automated systems for detecting hate speech on social media are built around machine learning models, which apply algorithms to training datasets to "learn" the patterns observed in real-world data. Off-the-shelf models of hate speech are likely trained on different contexts and definitions of hate speech, so may be too coarse to be used for most UN purposes, at least without substantial manual verification of result. As such, an automated classification system for a new use case will usually require the training of a new model.



[31] Ahmed, Z., Vidgen, B., & Hale, S. A. (2022). Tackling Racial Bias in Automated Online Hate Detection: Towards Fair and Accurate Detection of Hateful Users with Geometric Deep Learning. *EPJ Data Science*, 11(8). <https://doi.org/10.1140/epjds/s13688-022-00319-9>

In addition to labeled training examples, automated classification models also benefit greatly from having access to lexicons of specific terms commonly used in hate speech content. There are several publicly available sources for such data[30], in a wide range of languages, but it may also be worthwhile to develop an original dataset containing the most up-to-date and context-specific terminology seen in recent examples.

Detailed instructions for creating such a model are beyond the scope of this report, but we recommend that any such model (whether developed in-house or by a vendor) should meet the following requirements:

- ✔ The model should be specifically trained for this context and definition of hate speech and incorporate training data in all relevant languages and on all relevant topics.
- ✔ When pre-trained models (including word embeddings, LLMs, and off-the-shelf classifiers) are used as inputs, extra validation checks should be conducted to ensure that messages which are likely to be miscoded by pre-trained models (for example, because they use new or coded terms) are effectively handled by the final model.
- ✔ Model outputs should be delivered in the form of a predicted probability that a given message constitutes hate speech (and if applicable, hate speech of a specific type or level).
- ✔ The model's performance should be evaluated on a held-out set of labeled examples from the same context as its intended usage, with performance reported in terms of both its recall (*what percentage of online hate speech does it successfully catch?*) and its precision (*what percentage of messages classified as hate speech actually are?*).[32] When applicable, that performance should also be reported for different levels of hate speech and different targeted groups.
- ✔ Models should also be tested on real content by authors from various ethnic, religious, and/or political groups and across different ages and genders, if such data is available for the test set.
- ✔ All models should follow the guidelines set forth in the "Principles for the ethical use of artificial intelligence in the United Nations system" established in 2022. [33]

[32] In cases where training data is very limited, cross-validation can be used to evaluate performance on hold-out sets without excluding any training data from the final model.

[33] <https://unsceeb.org/principles-ethical-use-artificial-intelligence-united-nations-system>

All users of automated classification systems (including downstream recipients of data which have not been subject to human review) should be keenly aware of the limitations of these systems. A common aphorism in data science, attributed to George Box, is that “all models are wrong, but some are useful.” It should be taken as a given that all automated systems will make mistakes on a regular basis, both in flagging some benign content as hate speech and not flagging some actual hate speech. Given these limitations, the main uses of automated systems are (a) to screen potential online hate speech to narrow down the set of content requiring human evaluation (enabling more efficient review), and (b) to classify content for aggregate-level analyses in which some level of error is acceptable. Other uses should be approached with caution, and as a general rule, human review must be built into any process that involves responding to specific incidents of detected online hate speech.

Finally, based on our review of academic research on hate speech classification, we also recommend caution when applying off-the-shelf AI tools like ChatGPT and/or automatic translation services. As of this writing, those tools have shown significant potential, but their performance can be inconsistent, and their potential risks and limitations are not well understood. For example, the accuracy of automated translation tools varies greatly based on language, subject matter, and other characteristics, and in many cases they are insufficient for conveying critical information.[34] In the absence of other solutions, such tools may still offer advantages (for example, to enable flagging of potential hate speech in languages without labeled training data), but they generally do not perform as effectively as systems built for a specific purpose and context.

Classifying new messages

As new messages are posted on social media, they are then evaluated against the relevant definition of hate speech. Automated systems will do this either through real-time streams or batch processing, while manual systems will do so on a schedule that corresponds with reviewers’ work schedules. In either case, the key concern is timeliness: because social media can reach thousands or millions of people in a matter of minutes, detecting online hate speech quickly is crucial to effective responses.

When an instance of hate speech is detected, the next step will depend upon the intended usage. Determining the right workflow for this process is part of the planning work included in Step 1, and those involved in handling these messages should receive clear guidance and training on how to handle online hate speech when it is discovered in a way that aligns with the program’s mandate as well as broader UN policy.

[34] See for example <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8606479/>, which looked at the effectiveness of Google Translate for conveying medical instructions and found accuracy ranging from 94% to just 55% across seven languages tested.

STEP 4: DEPLOYMENT

The final step of implementation is to deploy this system as an ongoing process and integrating it with the overall program's workflows.

Setting up monitoring and alerts

To ensure proper response, a regular reporting process is required so that program staff can see when online hate speech is detected and act swiftly if required. This ideally involves some form of hate speech dashboard which displays key information in a user-friendly interface. Depending upon the specific data being collected and its intended usage, this information may include:

- ▶ Volume of online hate speech over time.
- ▶ Severity of observed hate speech.
- ▶ Specific examples of recent online hate speech from influential accounts or with high levels of engagement.
- ▶ Distribution of recent online hate speech (in terms of social media platforms used, known actors involved, targeted groups, and other key characteristics).
- ▶ Geographical distribution of online hate speech.
- ▶ Engagement with hate speech content (in terms of views, likes, shares, and so forth).
- ▶ Engagement with hate speech content (in terms of views, likes, shares, and so forth).
- ▶ Most influential users and groups responsible for online hate speech.
- ▶ Statistics on online hate speech content reported to platforms and resulting actions.

In less technologically-enabled programs, much of this information could be manually compiled using shared spreadsheets, periodic email reports, or similar approaches. The fundamental value of this deliverable, regardless of the specific form, is ensuring that those whose role it is to respond to online hate speech have easy access to the information needed to do so.

Depending upon the available responses, it may also be important to set up alerts (for example, using automated emails) that specifically flag the most noteworthy content and events.[35] Triggers for these alerts could include:



Incidents of online hate speech from especially high-profile individuals or groups (such as political or religious leaders).



Sudden spikes in online hate speech frequency overall or of a particular type (for example, in a specific area or targeting a specific person or group).



Incidents of online hate speech that surpass a threshold for very high engagement.

These alerts should mainly be used in circumstances where there is substantial benefit in minimizing delays between the event and the subsequent response, which most often occurs when the intended response is direct intervention with the actors involved or the social media platforms or the issuance of public statements. To avoid “alert fatigue” on the part of recipients, programs should be sure to set the alert criteria sufficiently high to avoid sending alerts too frequently or to an overly broad audience.

Maintaining and refining

Finally, online hate speech monitoring programs that are intended to persist for an extended period of time will need to be regularly improved in order to ensure their continued effectiveness. All steps in the process are subject to improvement, but the following actions are especially important for the long-term value of a program:

[35] These alerts are most effective when there are “normal” baseline levels of hate speech content to use for comparison, so when feasible it may be worth collecting data from wider timeframes and geographies than just those experiencing high levels of current hate speech.



Confirming or updating the program context and use cases to fit the evolving nature of a situation and/or the available responses.



Adding new platforms and forums which become relevant to online discussions and updating lists of influential accounts to monitor.



Updating keywords, hashtags, and other filter criteria to ensure comprehensive coverage of relevant content.



Adding new labeled training data for automated classification models, particularly for examples that are not well-handled currently, and retraining models with this new data.



Improving reporting and alerting systems to highlight the most actionable information based on feedback from users.

Assignment of these tasks to individuals should be done during the program planning phase, with improvements planned to take place on a weekly, monthly, or quarterly schedule as appropriate to the program.

SAMPLE PROGRAMS

A wide variety of programs are possible under this methodology, but as a starting point, we've developed three hypothetical examples below of how specific programs based on it might look. Their structure may be useful as a template for future programs, with parameters chosen and adaptations made to fit the new program's requirements. Each example addresses all the elements of each step shown in Figure 2 above, and follows the recommended guidance included in the preceding sections.

Example 1: The minimally-technical approach[36]

STEP 1: PLANNING

Anisa and Barbara are part of the UN country team in a host country that is about to hold national elections. Because past incidents of hate speech in the country have led to violence, they want to monitor incidents of potential online hate speech related to the elections.

To start, they decide to focus on election-related messages in the official national language, with a particular focus on messages from influential political figures that may target specific ethnic groups.

The intended usage of this data is to support advocacy to political party leaders where appropriate, which may include asking them to denounce specific incidents of online hate speech and discourage their followers from repeating or acting on them.

Because the intended response is advocacy rather than enforcement, they choose to use the broad definition of hate speech from the UN Strategy and Plan of Action on Hate Speech, because even less severe instances of hate speech can cause harm and should be discouraged.

STEP 2: DATA GATHERING

The team decides to manually collect data by creating accounts on each of the three most popular social media platforms in the country and following a list of the most prominent accounts from individuals and groups involved in the election.

Every morning, Anisa manually reviews these feeds and records specific instances of potential hate speech by copying the content and other information into a shared spreadsheet. She also takes screenshots of posts and saves them to a shared folder.

STEP 3: CONTENT CLASSIFICATION

After Anisa is finished reviewing the social media feeds and flagging potential hate speech, Barbara reviews each post to determine if it is indeed hate speech and, if so, its severity according to the UN definition.

STEP 4: DEPLOYMENT

At the end of each week, Barbara emails a report to the rest of the country team summarizing their findings from that week. It includes information on online hate speech patterns and examples of the most notable content, as well as a report on what actions were taken in response. Given a set of specific criteria, Barbara also alerts the team's Communications and Advocacy officer immediately when an especially severe or prominent incident of online hate speech is observed, to enable a quick response.

At the start of the next week, Anisa and Barbara meet to review how well this approach worked during the previous week, then decide what changes to make that week to improve the program.

[36] This example is for illustrative purpose only. UN responses to hate speech in and around elections will vary based on specific context.

Example 2: Leveraging social listening tools

STEP 1: PLANNING

Chike is a political affairs officer assigned to a UN peacekeeping mission in the wake of a conflict between two neighboring countries. While the leaders of those countries have signed a peace agreement, there are still incidents of localized violence, particularly against ethnic minorities in the border area.

These incidents are often preceded by online hate speech directed at these groups, and he suspects the violence may be a direct result of those messages. He decides to focus on content posted on the affected region by influential users that references ethnic groups or their leaders, in any of the three most common languages.

Chike intends to work with the four most popular social media platforms to remove content that constitutes incitement to further violence and has made connections with their trust and safety teams, who've agreed to partner on this effort.

Given this intended usage, Chike decides to focus specifically on hate speech that constitutes incitement to discrimination, hostility, and violence, since that speech is clearly prohibited by international law and subject to removal.

STEP 2: DATA GATHERING

To monitor speech across several different social media platforms, Chike obtains access to a commercial social listening platform that compiles posts from a variety of different platforms. He configures these tools to flag content from any user that reaches a minimum level of engagement and either references one of the affected ethnic groups by name or uses specific coded language often used to refer to those groups, in any of the three languages.

The social listening platform then monitors this content around the clock and presents a curated view Chike can access through its web interface.

STEP 3: CONTENT CLASSIFICATION

Twice a day, Chike reviews the content presented in the social listening platform interface, starting with the content that has the highest engagement, and evaluates it against the six-part threshold test for incitement in the Rabat Plan of Action. Because he is only fluent in two of the three languages, he uses automatic translation to initially screen content in the third language but brings in a colleague who is a native speaker to interpret any content that appears potentially to be hate speech. When a post is identified to constitute incitement, he applies a custom tag in the social listening platform to track it.

STEP 4: DEPLOYMENT

Using the social listening platform's analytics features, Chike sets up a dashboard to display online hate speech patterns and examples, which other colleagues from the mission are also able to use.

When new instances of online hate speech are identified, the platform also sends an automated message to the social media platform's trust and safety team for review.

Example 3: An automated early warning system

STEP 1: PLANNING

Dev is leading a team at UN Headquarters charged with preventing genocide and related crimes across the globe. Based on a survey of experts, they have identified 25 countries with a particularly high risk of such violence in the coming years. 8 of those countries already have substantial UN presences, so his challenge is to determine how to allocate his team's attention and resources across the remaining 17 countries.

He decides to use social media monitoring to improve his team's awareness of the situation in each of those countries. This system will monitor content across the most prevalent languages in each country and focus on online hate speech that is tied to race, ethnicity, religion, or nationality.

The goal of this program is to provide a high-level view of the social media climate in each country over time. When they see that hate speech in a given country is on the rise, they will then allocate their team's resources to investigate the local situation in depth and support to diffuse tensions that might otherwise lead to violence.

For this use case, all forms of hate speech are relevant, so Dev decides to apply the broader UN Strategy's definition of hate speech for overall tracking. If further classification of hate speech by severity is necessary, that can then be done by manual review by the relevant country team in accordance with the levels outlined in the [Detailed Guidance](#).

STEP 2: DATA GATHERING

For this large-scale program, Dev works with three of the most popular global social media platforms to get direct API access to posts from a wide range of highly influential accounts, and also to collect anonymized data on comments and replies to those posts.

His team then partners with a technology NGO to develop a data warehouse platform that processes and stores this data in a centralized repository.

Using samples of data collected on this platform, his team then engages local experts to create labeled training datasets for each country and language, with labels flagging online hate speech targeted at specific groups.

STEP 3: CONTENT CLASSIFICATION

The technology NGO partner team develops a workflow that automatically parses new content from each social media platform and produces structured data suitable for modeling.

They then use each country's training dataset to develop a multi-stage machine learning model that first classifies content as hate speech or not, then flags content that is likely directed at a particular group. These models are validated on held-out datasets of labeled content, with the error rates recorded for subsequent adjustment of aggregate estimates.

This process is run automatically around the clock, with the resulting classifications loaded into the data warehouse soon after the new posts become available.

STEP 4: DEPLOYMENT

Dev's team uses these classifications to produce several deliverables:

- A real-time map with trends across all 17 of the tracked countries, with an accompanying dashboard showing global trends.
- Real-time alerts for cases when a specific country shows an unusual spike in online hate speech, suggesting a rapidly developing situation that may require a more urgent response.
- An automated weekly report that goes into greater detail and highlights changes in patterns over the previous week, month, quarter, and year.
- Country-specific dashboards for the relevant country teams, to facilitate their own awareness and responses.

To ensure the system's long-term viability, the team engages with local experts and country teams on a quarterly basis to manually label new training examples, with a particular focus on content about which the existing model is uncertain. These experts also help them to update their lexicon of hate speech terminology in each country and language and highlight situations where new groups may need to be added to the list of potential online hate speech targets.

Dev's team also regularly checks in with the country teams to get feedback on potential improvements to the county-specific dashboards, learn about instances of incorrectly classified content, and identify new potential applications for their system.

Case Study: Costa Rica

Costa Rica's initiative to combat online hate speech represents a pioneering effort in the Americas, leveraging social media monitoring to address rising political polarization and hate speech. This program, in which the UN country team partnered with the national government, academia, private companies, civil society, media and other key sectors, involved a comprehensive strategy encompassing data collection, analysis, and multi-stakeholder engagement to mitigate online hate speech's impact. The initiative's success demonstrates the critical role of data-driven approaches and collaborative efforts in addressing social issues.

This case study gives an example of how the UN has effectively used social media monitoring for online hate speech in one country-specific context. Though this program did not have the benefit of a standard methodology to start from, it demonstrates how many aspects of this methodology look in practice, and the lessons learned in this program helped to inform the set of practices recommended in this report.

CONTEXT

In recent years, Costa Rica has experienced significant social and political polarization, and this situation became especially salient in the 2018 elections. This period saw an unprecedented rise in online hate speech, fueled by the emergence of an anti-immigration movement and other divisive issues. Recognizing the threat posed by unchecked online hate speech, UN Costa Rica embarked on developing a UN strategy, becoming the first country in the Americas to do so, and supported the government in launching a national strategy in 2024, the second country in the world to implement such a comprehensive approach.

PROGRAM DEVELOPMENT

The development of Costa Rica's online hate speech monitoring program was a structured and methodical process, largely mirroring the methodology introduced in this report. The program was led by the UN's country team, which began by securing support from the Office of the Special Adviser on the Prevention of Genocide (OSAPG) as well as partners in civil society and government. Its scope included the full range of speech that fell within the [UN Strategy and Plan of Action's](#) definition of hate speech.






The team started by gathering data to demonstrate the extent of online hate speech in Costa Rica. Through an alliance with the company Coes Communications, UN Costa Rica acquired access to the Brandwatch and Mention social listening tools to collect data from Facebook and Twitter (later adding Instagram and Reddit as well) and identified sets of keywords and hashtags to filter content down to the messages most likely to be hate speech. Flagged messages then underwent a manual review to confirm the accuracy of coding, followed by data graphing and analysis to develop a deeper understanding of the problem and identify potential responses.

RESULTS

The results were stark, revealing a 255% increase in online hate speech messages over three years. Specific targeted groups included journalists, women, migrants, LGBTQ individuals, and people of Asian descent. Looking at those who were spreading hate speech, the UN team found that the majority of online hate speech promoters were ordinary individuals rather than organized groups, with 12% of social media users identified as using hate speech of some form. Among this group, they found that adult men from urban areas were the most frequent promoters of online hate speech.

These findings led to the development of a comprehensive response strategy involving multiple stakeholders and creating resources to combat online hate speech legally and socially. To address the issue, the UN team formed alliances with various stakeholders, including the media, academia, and the national lawyer's association. They also developed advocacy materials and legal guides to address online hate speech. The culmination of this work was the government's adoption of the first national strategy on hate speech. This outcome emphasized the importance of data and methodology in mobilizing social and political will.

KEY TAKEAWAYS

-  The combination of commercial tools for data capture and a manual review process ensured the accuracy and relevance of the data collected, while enabling the team to monitor a much broader range of content than would be possible through a purely manual effort.
-  Building alliances with media, academia, and legal entities can amplify the impact of the initiatives.
-  Developing advocacy materials and legal guides helps in educating civil society and the public, providing clear frameworks for addressing online hate speech.
-  Balancing short-term results with long-term impact is crucial. Immediate successes help build momentum and sustain long-term efforts.
-  Generating social and political will is necessary for the success of any monitoring program. Data production is key to raising awareness and bringing together a broader movement to support initiatives.

Future Innovations

As online hate speech monitoring becomes a more widespread tool for preventing violence, hostility, and discrimination, UN entities and partners will encounter new opportunities and new challenges. In this section, we focus on several that came up repeatedly in our interviews and research.

SHARED INFRASTRUCTURE AND PROCESSES

While the general difficulty of data access was the most common challenge faced by online hate speech monitoring programs, a close second was the staff capacity technical skill required to ingest and analyze the massive volumes of social media content available. To overcome this challenge, these programs often rely on NGO partners, academic institutions, third-party vendors, and the platforms themselves to provide solutions. Yet even with this help, these programs are still often limited in their ability to make the most of the data available.

But while the particular characteristics of each program will vary, there are many common needs that can be met by shared tools and workflows. Many common aspects of the monitoring process could be generalized across programs, including:

- ▶▶ Gathering, processing, and storing social media content.
- ▶▶ Collecting examples of online hate speech from partners and the public.
- ▶▶ Facilitating manual hate speech classification.
- ▶▶ Training, deploying, and validating automated classification models.
- ▶▶ Analyzing patterns across time, geography, and other dimensions.
- ▶▶ Hosting dashboards and generating reports.
- ▶▶ Detecting anomalies and producing alerts.
- ▶▶ Due diligence, evaluation, and risk assessments.

Shared solutions to these needs would enable UN staff and partners to more efficiently and effectively implement new programs and ensure their long-term success.



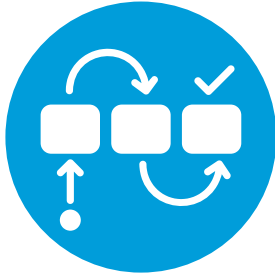
In practice, these systems would create reusable tools that are available for use by new programs, which could be deployed with minimal cost or delay. There are several precedents for this within the UN system already. Within the DPO, the Information Integrity Unit supports many of the department's peacekeeping missions by combining custom technology solutions, manual analysis, and specialized training and guidance to enable them

to effectively understand the social media environment. Meanwhile, the DPPA's Innovation Cell develops in-house tools for social media analysis (an example was the Sparrow tool, which generated automated reports using data from the Twitter / X API and delivered them to hundreds of users across the UN).

Not all shared solutions need to be developed in-house, however. One option would be to facilitate a shared procurement process for commercial software and services that could be used by multiple UN programs, to avoid the need for each program to evaluate and procure its own solutions. While the logistical challenges of managing the coordination and financing of such a procurement are significant and require careful contractual review, the

potential advantages include lower costs, cooperative training and best practice development, and reduced barriers to entry for new programs. A similar approach could be used in partnering with NGOs or academic institutions which would be able to provide ongoing capabilities in support of multiple UN programs, and potentially also other external partners.





Cooperative solutions could also pay dividends in terms of the workflows used for analysis and response. Much as some country teams coordinate their monitoring across the various departments represented, the expansion of online hate speech monitoring to more uses would be aided by improved coordination between entities to enable more effective and

timely responses, while leaving room for contextualized analysis and responses at national and local levels as required. In particular, the process for coordinating with social media platforms to address specific instances of hate speech that rise to the level of incitement could be made more consistent if there were a standard process that leverages standing, long-term relationships with the relevant platforms instead of ad hoc arrangements developed in the context of specific programs.[37]

IMPROVED VIOLENCE PREDICTION CAPABILITIES

As online hate speech monitoring becomes more common and consistent, it will generate more granular and reliable data to use for investigating the links between online hate speech and violence. This will enable research into many questions that can help to develop a more effective early warning system, including:

- ▶ Under what conditions is online hate speech most likely to lead to violence?
- ▶ What specific patterns and types of online hate speech are predictive of future violence?
- ▶ How do we assess the potential for specific hate speech to directly lead to violence?
- ▶ How and when can online hate speech interventions disrupt the cycle of violence?
- ▶ How can we evaluate online hate speech across countries to compare relative risk?





[37] A potentially useful model for this would be the Global Internet Forum to Counter Terrorism (<https://gifct.org>). Launched by major technology companies to coordinate their response to terrorism and violent extremism on their platforms, the group now works with various governments and civil society organizations to enable rapid reaction to prevent the proliferation of dangerous content. In a similar vein, a standing process for responding to incitement could enable more rapid and regular response to the most dangerous forms of hate speech.

All these questions have been the subject of ongoing research for years, but the introduction of consistent monitoring across many companies represents an opportunity for quantitative analysis at a scale which to date has never been impossible. Collaboration with academic researchers on these projects would further increase our collective knowledge. These discoveries could eventually lead to much better systems for assessing the risk of violence, and identifying effective responses, for a wide range of contexts across the globe - particularly emerging conflicts which may not yet have received the widespread attention of the international community.

THE BENEFITS AND CHALLENGES OF WIDESPREAD AI

To ensure the long-term viability of online hate speech monitoring, UN entities and partners will also have to adapt to the rapidly evolving capabilities of large language models and other AI systems. For monitoring online hate speech, these tools may eventually become sophisticated enough to facilitate automated detection with a precision comparable to what human coders can offer, even though such capabilities are decidedly beyond what is currently feasible. This would be particularly impactful in situations where the volume of content to monitor greatly exceeds our ability to process it all on a timely basis, especially if these tools become much more adept at interpreting non-English text and processing non-text content (such as audio and video).

At the same time, the widespread availability of advanced AI tools creates new and daunting challenges for monitoring and addressing online hate speech, requiring additional human oversight from trained experts capable of recognizing and countering such automated systems. Some of these potential complications include:

-  Automated generation and amplification of online hate speech content, leading to much greater volumes and reach than is possible by hand.
-  Adaptation of online hate speech content, including generation of unique images and video, in order to circumvent platform moderation tools.
-  Proliferation of deep fakes and manipulated media to give false credibility to online hate speech content.
-  Increased potential for outside actors, including international interests, to promote hostility and division by using AI to impersonate local leaders, groups, or individuals.

Artificial intelligence has quickly escalated the competition between the promoters and opponents of hate speech, as its growth offers new capabilities that can be used for good or ill in similar amounts. Just as social media platforms are forced to improve their moderation tools in response, online hate speech monitoring programs must continue to leverage the most advanced tools available to find and address online hate speech to limit its damage. This situation only amplifies the need for collaboration and innovation across the international community, to ensure that our efforts to prevent violence, hostility, and discrimination keep pace with the fast-changing technological circumstances.

Conclusion

This methodology represents a significant advance in addressing online hate speech globally. It provides a standardized approach that integrates best practices from across various sectors to better understand and combat the potential real-world impacts of online hate speech, including violence and genocide. This report also makes clear the importance of collaboration with tech companies, NGOs, and other partners to improve access to social media data and develop more effective monitoring tools. Forthcoming technological advancements, particularly in the areas of AI and machine learning, present new opportunities for accurately detecting and classifying hate speech across diverse languages and contexts. But these same innovations also create new challenges, as they are likely to amplify the volume and variety of online hate speech, making it more difficult to identify and respond to.

To effectively monitor and address online hate speech in the years to come, it is essential to focus on training UN staff and partners, continuously evaluate and improve this methodology, and use monitoring data to influence policy and advocacy efforts. These steps will ensure that monitoring programs not only contribute effectively to preventing hate speech-related violence but do so in a way that protects human rights and adheres to ethical standards and guidelines. Ultimately, the success of this methodology will depend on cooperative efforts across the UN system to mitigate the risks associated with online hate speech and prevent future acts of violence and discrimination.

Appendix A

The United Nations Strategy and Plan of Action on Hate Speech

<p>The main website for the UN Strategy and Plan of Action on Hate Speech can be found at</p>	<p>https://www.un.org/en/hate-speech/un-strategy-and-plan-of-action-on-hate-speech.</p>
<p>A PDF version of the full document is available at</p>	<p>https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf.</p>
<p>Detailed guidance on implementation for UN field presences is available at</p>	<p>https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20PoA%20on%20Hate%20Speech_Guidance%20on%20Addressing%20in%20field.pdf.</p>

Appendix B

Glossary of Key Technical Terms

API (Application Programming Interface):	A service that provides an automated way to access data through a programmatic request - for example, by requesting all social media posts that match a given keyword and timeframe.
Coded Speech:	Language which uses seemingly benign terms to refer to targets of hate speech without mentioning them explicitly, in order to avoid detection and censorship
Content Moderation:	The process of monitoring and managing user-generated content based on platform-specific rules and guidelines to prevent harmful or illegal content.
Crowdsourcing:	The practice of obtaining information or input into a task or project by enlisting the services of a large number of people, typically via the Internet.
Data Parsing:	Analyzing a collection of data in a particular format (for example, a text message or image) to extract useful information.
Engagement Metrics:	Data points that measure how users interact with content, including views, likes, shares, and comments.
Graph Data:	Information about the network of relationships between entities (like social network connections between people).

Keyword Filtering:	Using specific words and phrases to filter content automatically, often to identify or block certain types of content.
Lexicons:	Collections of words and phrases associated with a specific language or subject matter.
Machine Learning:	A branch of artificial intelligence that involves training a computer system to learn from data, recognize patterns, and make decisions with minimal human intervention.
Metadata:	Data that provides information about other data, such as the source of a social media post, time of posting, and device used.
Natural Language Processing (NLP):	The use of computers to process and analyze large amounts of natural language data.
Sentiment Analysis:	The process of computationally determining whether a piece of writing is positive, negative, or neutral.
Social Listening Tools:	Tools used to monitor and analyze online conversations on social media platforms to gain insights into users' opinions and behaviors.

Social Media Platforms:	Online services accessed through a website or application that enable users to create, share, and consume content such as text, images, audio, or video.
Training Datasets:	Sets of data used to train machine learning models to recognize patterns and make decisions.

Appendix C

Assessment of Existing Research and Tools

This section reviews the latest research into online hate speech, examines the methodologies used in this work, and discusses the relevant tools and data available to UN programs and partners.

SUMMARY OF PREVIOUS RESEARCH FINDINGS

The history of online hate speech monitoring can be traced back to the early days of the internet, when forums and chat rooms were the primary spaces for digital interaction. Initially, the monitoring was manual, relying on reports from users and the discretion of moderators to identify hate speech. As online platforms grew, the scale of the problem became apparent, necessitating more sophisticated approaches. The development and deployment of automated systems for detecting hate speech have been central to this evolution, and the sophistication of approaches used for this purpose has kept pace with the broader advancements in computer science over the past decades.

Researchers in the field have also faced numerous challenges. One significant issue has been the definition of hate speech itself, which varies widely across legal jurisdictions, cultures, and media organizations, making universal standards for detection difficult to establish.[38] Additionally, the linguistic subtlety of hate speech, which can involve sarcasm, coded language, and cultural references, makes automatic detection challenging.[39] Data scarcity, especially in languages other than English, and the lack of labeled datasets for training machine learning models, have also been significant obstacles. And recent restrictions on research access to social media data, due to both platform changes and data privacy regulations, have hit particularly hard on a field that was built on easy access to text-based examples from Twitter.

Extending this research outside of the academic context provides additional issues, which affect both civil society actors and the technology companies that try to reign in hate speech. The adaptability of hate speech, with perpetrators constantly evolving their language and methods to evade detection, means that the problem of detecting such speech can never be conclusively solved. The effectiveness of these technologies in diverse cultural and linguistic contexts is also quite limited, with the latest AI tools heavily focused on English-language content. More generally, the rapid evolution of online language and the emergence of new platforms complicate efforts to keep detection methods current, so staying on top of the latest developments requires continual investment and innovation.

[38] Coalition for Independent Technology Research, 2023

[39] Bigoulaeva et al., 2023

ANALYSIS OF HATE SPEECH IDENTIFICATION METHODS

Early systems for identifying hate speech relied on simple keyword-based filters, but these were quickly found to be both over- and under-inclusive, missing nuanced or coded language while also flagging benign content. Subsequent researchers improved on these techniques through the use of machine learning models, which enabled probabilistic classification of content instead of all-or-nothing rules. These algorithms became much more effective when combined with natural language processing (NLP) tools such as pre-trained word embeddings, named entity recognition, and shared lexicons of key terms, which enabled models to bring in much more information than just the words used in any given training dataset.

All of these approaches rely on a purpose-built classification system built using traditional machine learning workflows. But with the development and widespread availability of large language models such as ChatGPT in recent years, some researchers have wondered whether these general-purpose models could be utilized to identify hate speech effectively. So far, the results have been underwhelming. While these tools can be effective at identifying hate speech (particularly with carefully designed prompting strategies), their performance at hate speech classification has been comparable to that seen with previous models.

More recently, the introduction of transformer-based models represents a significant leap forward in the ability to understand and interpret the complex nuances of language used in online hate speech. These models can consider the context and subtleties of language, offering improved detection rates. The development of cross-lingual transfer learning methods has also expanded the reach of monitoring efforts to languages with fewer resources. Additionally, the consolidation and analysis of large datasets across multiple platforms have enabled more comprehensive studies on hate speech patterns and the effectiveness of detection mechanisms.

These methods have shown reasonably good accuracy at identifying hate speech by machine learning standards, but never sufficiently high to remove the need for human-in-the-loop verification before action is taken on any individual piece of content. [40] Moreover, this performance varies greatly across contexts and languages, with detection in non-English languages (which would almost certainly be required in most of the UN's use cases) a much more difficult challenge. These methods also depend on having access to hand-coded training data for a given context, which would need regular updating as situations and language evolve, so a one-size-fits-all solution is not realistic in the foreseeable future.

[40] Specific accuracy rates vary widely and depend upon the frequency of hate speech in the test dataset, the definition of hate speech used, and the type of content being classified. But as a general indicator, very few models have been able to show greater than 90% accuracy, meaning that they will incorrectly classify at least 1 out of 10 examples where the correct classification is not obvious.

AVAILABLE TOOLS AND DATA

This section lists a variety of available tools and data sources for social media monitoring and provides basic information about their potential capabilities and usage. These tools range from in-house tools developed at the UN to paid tools available for purchase. For commercial offerings, while it would be inappropriate to endorse any specific products, we have chosen to mention tools which were identified in our research for this project as having been successfully used by UN entities or other organizations for hate speech monitoring and/or related purposes. Choosing between these options (as well as others not listed here) should be done through a more detailed evaluation during the planning phase of each program.

Existing tools and services

From UN entities:



The DPPA-PMD Innovation Cell offers research and analysis of social media dynamics, including on hate speech. It also scans periodically for emerging technologies to contribute to DPPA's analysis and early warning efforts and hosts the annual E-Analytics and Innovation course.



UNDP's eMonitor+ is an open-source technology which provides a secure environment for collaborative fact-checking to verify the accuracy and authenticity of digital content. This AI-powered system is also trained to map digital space and monitor trends on various topics, such as misinformation, hate speech, political polarization, and online violence against women, journalists, and marginalized groups. This provides the foundation for informed, evidence-based strategies and works towards creating safer and more inclusive digital environments. Available in five languages (Arabic, English, French, Spanish and Portuguese), over two million pieces of online content have been mapped with eMonitor+.



iVerify is UNDP's fact-checking tool to identify false information, prevent and mitigate its spread. As a digital public good, iVerify provides national stakeholders with a support package to enhance identification, monitoring and response capacity to threats to information integrity. The support package includes digital tools, capacity building modules, partnership opportunities, and communication and outreach strategies amongst others.



For missions within the DPO, the department's Information Integrity Unit offers specialized training, support, and tooling to monitor misinformation, disinformation, and hate speech. These offerings are designed to offer a higher level of technical sophistication than individual missions can develop on their own, including leveraging integrations with platforms including Facebook, Instagram, Twitter/X, YouTube, and Telegram.



Together with DPO's Information Integrity Unit, OICT's Enterprise Solutions Service is rolling out Unite Wave, a machine learning-based tool for monitoring, transcribing and analyzing online and offline radio broadcast data as well as transcriptions from videos such as those on YouTube. The tool can transcribe speech in over 70 languages, with more being added.

From social media platforms:

CrowdTangle

➤ CrowdTangle is a social media analytics tool available from Meta which aggregates data from across the company's Facebook and Instagram platforms. Provides a curated view of public content from the most prominent accounts, with a web-based user interface showing what content is being posted and by whom, and the associated engagement metrics such as likes, comments, and shares. Users can sort through content based on search terms and filters and track trends over time.

➤ In March, Meta announced that CrowdTangle would be discontinued as of August 2024.

Meta Content Library

- ▶ Successor to CrowdTangle, providing much of the same functionality but with a more extensive underlying dataset.
- ▶ Adds the ability to directly analyze data in Python and R as well as through the web interface, albeit in a sandboxed environment with limited ability to export results.
- ▶ Data is hosted by the University of Michigan’s ICPSR, which manages access.
- ▶ As of March 2024, it is still being rolled out and was not available for testing for this report.

From commercial software providers, by category:

Hate speech monitoring platforms: These tools are purpose-built for detecting hate speech online and could potentially provide an end-to-end monitoring solution.

- ▶ Nisien (<https://nisien.ai>)
- ▶ TrustLab (<https://www.trustlab.com>)

Social listening platforms: These tools all compile social media data from across multiple platforms and enable the user to track trends and view content through a single web interface. They are most commonly used for consumer intelligence and brand management, but can readily be adapted to monitor other kinds of content by selecting relevant filtering criteria. They differ in the data sources available and the analytic capabilities included, particularly in terms of advanced content classification and analysis, and have varying degrees of applicability to civic use cases.[41]

- ▶ Meltwater (<https://www.meltwater.com>)
- ▶ Brandwatch (<https://www.brandwatch.com>)
- ▶ TalkWalker (<https://www.talkwalker.com>)
- ▶ Synthesio (<https://www.synthesio.com>)
- ▶ Digimind (<https://www.digimind.com>)
- ▶ Mention (<https://mention.com/>)

[41] For more information about available social listening tools and an analysis of their strengths and weaknesses, see the National Democratic Institute’s report, “The Changing Landscape of Social Media Monitoring Tools”, available at <https://www.ndi.org/publications/changing-landscape-social-media-monitoring-tools>.

Digital intelligence platforms: These services are designed to analyze social media data for harmful content such as misinformation, disinformation, and threats, and could potentially be adapted to fit the hate speech use case. They generally include much better content classification capabilities than social listening tools and would be more effective at tracking quantitative trends as a result.

- ▶▶ Omelas (<https://www.omelas.io>)
- ▶▶ Logically (<https://www.logically.ai>)
- ▶▶ Recorded Future (<https://www.recordedfuture.com>)
- ▶▶ Alethea (<https://www.alethea.com>)

In addition to the “off-the-shelf” tools listed here (which may require some customization, but largely fit the use case already), there is a wide variety of other social media analysis and monitoring tools available that could be adapted to facilitate hate speech monitoring. Implementing a program that uses these tools would require partnering with developers to add new capabilities, such as automated hate speech classification. This would require a substantial investment of resources but could be worth considering for situations where existing tools’ functionality is insufficient, particularly for programs which are expected to persist for an extended timeframe.

When procuring commercial or other third-party tools, programs must also ensure that their intended usage complies with the acceptable use policies of the developers. The specific nature of the intended usage should ideally be noted in the applicable contract or purchase order, to avoid any unwelcome surprises after the program has begun.

Data sources

Social media data:

Platform APIs: These services allow users to directly access a selection of content posted on their platforms through data extraction pipelines or other automated systems. Platforms that offer these APIs include:

- ▶▶ Discord (<https://discord.com/developers/docs/intro>)
 - ▶ Available at no cost.
 - ▶ Enables fully automated interaction with the Discord platform, comparable to what a user could do through the platform itself, with a dedicated real-time API for live events.

- Rate limits built into Discord API.
- Requires the use of one of a wide range of third-party tools for interaction, and availability of data may vary across independently hosted Discord servers.

➤ Reddit (<https://www.reddit.com/dev/api>)

- Limited API available for free to all, research access available upon request.
- Allows access to publicly available content organized by user, forum, and other criteria.
- Rate limits enforced based on user type.

➤ Telegram (<https://core.telegram.org/api>)

- Available at no cost.
- Allows access to all content and activity available through the Telegram platform to a specific user.
- No limits on data quantity, but does not provide historical data access or access to private channels.

➤ Twitter/X (<https://developer.twitter.com>)

- Available for a fee, though some UN entities have access through an existing institutional agreement.
- Allows selection of content data through keyword-based and user-based filtering.
- Data limitations vary based on the type of agreement, with publicly listed pricing for subscription tiers ranging from \$100 USD per month to \$500,000 USD per year.

➤ YouTube (<https://developers.google.com/youtube/v3>)

- Available at no cost.
- Allows access to videos, comments, captions, and more from specific accounts and through a search feature.

- Access is subject to a standard quota, with the ability to request quota increases on a case-by-case basis.

➤ As of this writing, no directly accessible APIs could be found for Facebook, Instagram, or TikTok data. (TikTok does offer a research API for approved academic researchers but does not make it available to other kinds of organizations.)

Third-party bulk datasets: Various sites offer access to historical social media datasets from platforms such as Twitter and Reddit, but our research did not identify any such resources which were conducive to ongoing monitoring of new content. (Social listening platforms collect this kind of data internally, but generally do not provide access other than through their proprietary tools.)

Hate speech training and vocabulary data:

Labeled hate speech examples: These datasets included manually labeled examples of hate speech content (including non-hate content for comparison), mostly from published data used in academic research. They vary widely in the specific definitions of hate speech used, the types of hate speech included, and the contexts from which the data was obtained.

➤ Hate Speech Data (<https://hatespeechdata.com>)

- Extensive list of hate speech datasets and keywords in 25 languages.
- Primarily taken from academic research papers.

➤ MetaHate dataset (<https://arxiv.org/html/2401.06526v1>)

- A standardized dataset of English-language hate speech content aggregated from 36 separate academic datasets.
- Includes 1.2 million social media posts taken from various platforms.

Lexicons of hate speech vocabulary: These datasets include information on the specific language used in incidents of hate speech, which can be used to determine appropriate filtering keywords and to improve hate speech classification models.

▶ Hate Speech Data (<https://hatespeechdata.com>)

- ▶ Extensive list of hate speech datasets and keywords in 25 languages.
- ▶ Primarily taken from academic research papers.

▶ The Weaponized Word (<https://weaponizedword.org>)

- ▶ A licensable tool that extends of the open-source Hatebase project, which shut down in 2022.
- ▶ Includes a 7,500+ word index of hate terminology in more than 100 languages, which is continually expanded by local language experts and can be further developed for specific use cases.
- ▶ Offers API access to various specialized lexicons and other tools on a subscription basis.