TABLE OF CONTENTS

INTRODUCTION **2**

SCOPE **2**

CHALLENGES IN DEFINING HATE SPEECH **2**

WHO IS THE AUTHORITY ON HATE SPEECH THEN? **4**

DEFINING ONLINE VIOLENCE **4**

TROLL FARMS AND INAUTHENTIC BEHAVIOURS **6**

MONITORING ONLINE CONTENT **6**

DIGITAL THREAT IDENTIFIER 7
PREPARING YOUR SOCIAL MEDIA ACCOUNT 7
IDENTIFYING KEYWORDS 8

CIRCUMVENTING CONTENT MODERATION **9**

PRIVACY AND SAFETY OF RESEARCHERS **10**

TRACKING ONLINE CONTENTS **11**

DATA COLLECTION 11
STEPS TO COLLECTING DATA 11
DATA ANALYTICS, LISTENING AND TRACKING TOOLS 13
CrowdTangle *13*
Talkwalker *16*
Keyhole *18*
NewsWhip *18*
Tableau *20*
Brandwatch *21*
ADVOCACY 22

USEFUL RESOURCES **23**

REFERENCES **24**

Introduction

Even though social media plays a critical role in facilitating access to information while empowering ordinary citizens to produce contents that matter to them, it is also used to fuel hate, conflict, polarization and disinformation [8]. Identifying these digital threats and protecting users from online abuse have drawn the attention of researchers, advocates, policymakers, and social media companies. However, the response by social media companies whose very platforms are used to spread such content has been complicated by the lack of accord on the boundaries of acceptable online conduct [9].

Reliably mapping digital threats on social media platforms, hence, requires a distinct strategy in the monitoring and documentation of harmful online content. As most of GIF's partners [10] are human rights defenders, civil society actors and journalists from close to 50 countries who advocate for an inclusive, safe, and open Internet, this methodology is framed to accommodate the respective needs and end-goals of this diverse partnership.

While this methodology document seeks to provide partners with the tools to identify, document and report harmful speech, it is expected to eventually lead to a large entry of important datasets that could potentially indicate the most suitable line of action to counter online harmful speech. Additionally, the interrelationships between these harmful sentiments and their corresponding offline harms can be accurately analyzed, then used by partners to promote effective local and transnational solutions.

In developing this methodology document, GIF analyzed relevant studies that have been conducted on the topic and found that there is generally one main approach for monitoring and documenting harmful content online. The approach entails identifying and cataloguing keywords and the user accounts of bad actors onto a versatile template, while providing historical, linguistic as well as other relevant context about each item of relevant content. This approach can be adjusted to meet the specific needs of the researcher, depending on the prevalence of the online harm and its offline manifestation.

Scope

This document presents a digital threat tracking and mapping methodology that aims to enable external partners to aptly identify, track and report harmful social media contents. However, understanding the definitions of the terms and expressions pertinent to digital threats and building a rich context thereof should precede tracking and mapping activities.

The key justification for studying these definitions is that understanding the cultural and linguistic nuances of these terms and expressions within the online space in question plays an exclusive role in reliably mapping the social media landscape where groups and communities are being targeted.

Cognizant of its expansive domain, therefore, a decision was made to develop this methodology based on the two most important topics of online harmful practices – hate speech and online violence. Because limiting the purview of this document would enhance its potency, it focuses on the three social media platforms that a majority of online users choose to access digital news contents from – Facebook, Twitter and YouTube [29].

Accordingly, from the many forms of digital threats, this methodology focuses on how partners can build a comprehensive knowledge, understanding and analysis of the impact of hate speech and online violence within their respective regions so that an effective data-driven advocacy can be crafted. However, depending on specific needs, this methodology can also be adapted to address other forms of harmful practices.

Challenges in Defining Hate Speech

To understand how experts characterize hate speech and to acquire cumulative interpretations of hate speech and its impacts, it is important to look at the most notable terminologies and how they are defined in academic, legal and policy literature. Ideally, having an unequivocal definition of hate speech would allow for a coherent way of dealing

with its online and offline manifestations, which can also be used to address specific problems with the right contextual awareness.

Seeking to produce a "legal literature" that aimed to define hate speech and hate crime, a group of researchers conducted a systematic review on the evolution and status of hate speech [27]. The group reviewed research papers that were written in English and Spanish, which studied hate speech from legal and communication standpoints.

According to the review, academic interest in hate speech grew from 42 to 162 in five years (2013 – 2018), measured by the volume of indexed papers produced on the Web of Science, the world's first citation index. Unfortunately, hate speech remains one of the most important topics that lacks uniform legal definition across countries and institutions.

According to the United Nations, for example, there is no universal definition [24] of hate speech under the *International Human Rights Law* because the concept is still disputed based on freedom of opinion and expression.

Nevertheless, the UN provides a definition under a framework known as *United Nations Strategy and Plan of Action*, which defines hate speech as:

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

The UN elaborates that hate speech can be expressed through images, cartoons, memes, objects, gestures and symbols, which often generate intolerance, hatred and, in certain contexts, can be demeaning and divisive. It notes that these expressions are not considered hate speech if they are directed against States and their offices or symbols, public officials, or religious leaders or tenets of faith.

On the other hand, the United Nations Educational, Scientific and Cultural Organization (UNESCO) [34] defines hate speech even more cautiously. In a study it developed in cooperation with the Programme in Comparative Media Law and Policy (PCMLP) at the University of Oxford, UNESCO insists that any collateral restriction of legitimate expression in response to and while identifying and limiting hate speech would not be justifiable. Nevertheless, this study states that speech can be identified as hate speech (whether it is expressed through text, image, or sound) by its degrading and dehumanizing purpose against members assigned to a targeted group.

Similarly, the Office of the High Commissioner for Human Rights (OHCHR) [25] held a series of expert seminars between 2008 and 2012 to deliberate on the prohibition of incitement to national, racial or religious hatred, in which it adopted the Rabat Plan of Action, a document it called a "practical and useful tool" in the fight against incitement to hatred and violence. The Rabat Plan of Action, which has since been translated in 32 languages, outlines a six-part threshold test to identify hate speech, namely:

1. Context: The social and political context of the speech,
2. Speaker: Status of the speaker,
3. Intent: the intent to incite the audience against a target group,
4. Content and form: the content and form of the speech,
5. Extent: Extent of the speech's dissemination, reach and magnitude,
6. Likelihood and imminence: the likelihood of harm, including imminence.

The action plan suggests that all of six parts of the threshold test need to be fulfilled for a statement to be considered as an offence so that the limitations to freedom of expression remain an exception.

In line with the UN and its specialized agencies, scholarly journals and researchers also exhibit identical caution when it comes to defining hate speech. In its mapping review of approaches to defining and measuring hate speech, a journal by Campbell Systematic Reviews [35] found that there is a rapid increase of vocabulary used to refer to similar harmful online behaviours, which, together with the lack of clarity in relevant definitions, pose real challenges to policy, practice and research.

Who is the Authority on Hate Speech then?

Amid the challenges caused by nuanced and pronounced definition differences, the question remains, who is the arbiter for determining hate speech? While one of the main reasons all the institutions defining hate speech found the task to be complex is its delicate intersection with freedom of expression, another practical issue is the lack of jurisdictional control over online content. The increasing speed with which content spreads on the Internet makes it challenging for state actors to execute national laws on online platforms. In some instances, governments have taken regulatory measures [15] that pressed tech companies to take more responsibility in enforcing relevant laws.

However, such measures are limited in their applicability, which has so far allowed social media platforms' community standards to largely remain the governing principles regarding user practices. Ironically, because social media companies continue to rely on users to flag problematic content such as hate speech, users have essentially become the authority who have the last word. Hence, the role of the media, human rights advocates and local institutions becomes irreplaceable in raising public awareness about the roles and responsibilities they hold in making the virtual world free and safe. Therefore, it is essential to outline the community standards provided by some of the popular social media platforms.

META

Instagram's community guidelines states that "it is never OK to encourage violence or attack anyone based on their race, ethnicity, national origin, sex, gender, gender identity, sexual orientation, religious affiliation, disabilities or diseases." On the other hand, Facebook's community standards [13] state that people will freely connect if they do not feel attacked based on who they are. And to that effect, it defines hate speech as a direct attack against people based on what it calls protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. By 'attack' it is referring to violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. It also prohibits harmful stereotypes, which it describes as dehumanizing comparisons that have historically been used to attack, intimidate, or exclude specific groups, and that are often linked with offline violence.

TWITTER

Twitter's [31] policy on hateful content, similarly, prohibits the promotion of violence against or direct attacks on or threats to other people based on race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. Its policy indicates that women, people of color, lesbian, gay, bisexual, transgender, queer, intersex, asexual individuals, marginalized and historically underrepresented communities are disproportionately targeted with abuse online. The policy expresses its dedication to fighting abuse that is inspired by hatred, prejudice or intolerance, especially abuse that aims to stifle the voices of those who have historically been marginalized.

YOUTUBE

The hate speech Community Guidelines by YouTube [14] prohibit content that promotes violence or hatred against individuals or groups based on their age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their kin and veteran status. It also does not allow content that encourages violence against individuals or groups based on any of the listed attributes. Additionally, its guidelines state that threats and implied calls for violence are forbidden.

Defining Online Violence

The United Nations Population Fund (UNFPA) [1] uses the term digital violence to describe a similar abuse, though it admits there are many other expressions for it. According to a study it attributed to the Economist Intelligence Unit, 85% of women globally have been exposed to or experienced some form of online violence, and nearly 40% have experienced it personally. The UN agency defines online violence as "an act of violence perpetrated by one or more individuals that is rooted in gender inequality and gendered norms and that is committed, assisted, aggravated and amplified in part or fully by the use of information and communication technologies or digital media, against a person on the basis of their gender."

An online community hub, Tech Without Violence [5], used a different terminology for a similar concept – cyberviolence. It defines cyberviolence as "online behaviours that criminally or non-criminally assault, or can lead to assault, of a person's physical, psychological or emotional well-being."

The European Council [4], on the other hand, used cyberviolence in slightly broader context. The council defines it as "the use of computer systems to cause, facilitate, or threaten violence against individuals, that results in (or is likely to result in) physical, sexual, psychological or economic harm or suffering and may include the exploitation of the individual's circumstance, characteristics or vulnerabilities."

As indicated above, the user policies of respective social media platforms are by far the main governing principles regarding user practices. While it is important to study the different measurements that institutions use to define hate speech, ultimately the presiding standard is that of the platforms.

META

Meta's policy [36] puts incitement and violence into one category and defines it as an abuse that has a potential offline harm. These abuses are threats that could lead to death (and other forms of high-severity violence) and admission of past violence targeting people or places. The community standard further goes to define 'threat' as follows:

- Statements of intent to commit high-severity violence,
- Calls for high-severity violence,
- Statements advocating for high-severity violence,
- Aspirational or conditional statements to commit high-severity violence,
- Statements admitting to committing high-severity violence.

TWITTER

Twitter categorises violence into two: *Violent Threats* [32] and *Glorification of Violence* [26] *acts*. It defines violent threats as "statements of an intent to kill or inflict serious physical harm on a specific person or group of people. Threatening an individual or a group of people with violence as well as the glorification of violence is prohibited on the platform. Others in this policy include, but are not limited to:

- Threatening to kill someone,
- Threatening to sexually assault someone,
- Threatening to seriously hurt or lead to someone's death or serious physical injury,
- Asking for or offering a financial reward in exchange for inflicting violence on a specific person or group of people,
- Glorifying, celebrating, praising or condoning violent crimes, violent events where people were targeted because of their membership in a protected group, or the perpetrators of such acts.

The second category – glorifying violent acts – is contents that "glorify acts of violence in a way that may inspire others to replicate those violent acts and cause real offline harm, or events where members of a protected group were the primary targets or victims." According to the platform, glorification is praising, celebrating, or condoning statements towards:

1. Violent acts committed by civilians that resulted in death or serious physical injury,
2. Attacks carried out by terrorist organizations or violent extremist groups, and
3. Violent events that targeted protected groups

YOUTUBE

Content that seeks to "shock or disgust viewers, or content that encourages others to commit violent acts, are not allowed on YouTube" [37]. Below are some of types of content that can be categorized as violent behaviour:

- Inciting others to commit violent acts against individuals or a defined group of people.
- Fights involving minors.
- Footage, audio, or imagery involving road accidents, natural disasters, war aftermath, terrorist attack aftermath, street fights, physical attacks, immolation, torture, corpses, protests or riots, robberies, medical procedures, or other such scenarios with the intent to shock or disgust viewers.

- Footage or imagery showing bodily fluids, such as blood or vomit, with the intent to shock or disgust viewers.
- Footage of corpses with massive injuries, such as severed limbs.

Additionally, YouTube does not allow violent or criminal organizations to use the platform "for any purpose, including recruitment." Users cannot use its platform to praise, promote, or aid these organizations.

Troll Farms and Inauthentic Behaviours

With the advancement of artificial intelligence, social media is no longer reserved for only human users. Studies suggest that bots and automated accounts are increasingly being used to shape public opinion. [21]

UNDP published a research paper comparing organizations based in 28 countries, in which it found that these organizations had deployed cyber troops to operate fake accounts that flooded online platforms with spam and fake content. [2]

"Many different countries employ significant numbers of people and resources to manage and manipulate public opinion online, sometimes targeting domestic audiences and sometimes targeting foreign publics," the paper adds, while explaining the omnipresent nature of cyber troops. These troops, backed by trolls and bots, produce content on fake social media accounts to artificially amplify the reach and influence of content that is mostly harmful.

Similarly, according to the North Atlantic Treaty Organization (NATO), these entities, called troll factories, often conceal their identity to "achieve their goals using, among other things, fake news and hate speech." [23] The term 'troll factory' began to gain traction in 2015, NATO's modest brochure affirms.

"Since the troll factory employs hundreds of people and each employee has several accounts, it is easy to create a social network linking fake profiles and the impression of a real one. The longer the given accounts are maintained, the easier it is to create the illusion of reality," it adds.

Another term that Meta created to describe inauthenticity is CIB, short for Coordinated Inauthentic Behaviour. Meta defines such accounts as "groups of pages or people to mislead others about who they are or what they're doing." [16] According to the platform, these behaviours rely on fake accounts and other inauthentic tactics to drive traffic and mislead people while most of the content promotes dangerous, harmful, or hateful content.

The problem, Meta adds, "is that the actors behind these campaigns are using deceptive behaviours to conceal the identity of the organization behind a campaign, making the organization or its activity appear more popular or trustworthy than it is, or evade our enforcement efforts."

Monitoring Online Content

Analysing raw data is only as good as the data itself. Patchy, imprecise, or inconsistent data could lead to an analysis that is incomplete, or worse, misleading. Below is a list of tips that would serve as an initial starting point to systematically approach the tracking and documentation of online content. It will help practitioners to develop datasets and benchmarks to analyze and contextualize online harms as well as to indicate ways that prevent offline harms caused by online content.

To that end, this methodology studied a few initiatives that had impactful monitoring and documentation works across the world. In India, Equality Labs [28] published a report that discusses the threats against Indian minorities in India and abroad. The report provided essential data, analysis and recommendations following a six-month research project that monitored the prevalence of hate speech against minorities based on caste, religion, and gender on Facebook.

In Kenya, the Umati [33] project set out to monitor online content and record incidents of hate and dangerous speech after it found out there is a need to define, identify and deal with dangerous speech during the 2013 elections. According to the report it published immediately after the elections, the project sought to identify and understand the use of dangerous speech online, which recommended non-government solutions to the offline harms it created.

This methodology aims to expand on these extensive and focused research projects so that a comprehensive monitoring, identification and reporting tool is available to support future researches.

Digital threat identifier

Monitoring and identifying digital threats require a sustained effort so that accurate and reliable datasets are built. Furthermore, investigating these threats should be done in a way that ensures accurate results within an environment that is both safe and comfortable for the researcher. Having a dedicated social media account for such tasks is, therefore, advisable so that partners can stay away from unwarranted exposures and incidents.

The other reason to having a separate account is that of the personalized algorithms that social media platforms implement on news feeds to improve their user experience, whose parameters are, by and large, criticized for their biases. A case study [11] done in 2018 on the Italian elections found that Facebook's News Feed algorithms tend to reinforce "likes" by filtering and favouring those posts among the available contents. The study concluded that this bias is 'non-negligible' even for users who have a 'neutral' political view. And to avoid the limitation caused by such types of platform-imposed biases, insulating accounts pertinent to this task is advantageous.

According to a cross-national perspective provided in the book, *Online Hate and Harmful Content*, hate, motivated by prejudice, can be expressed in speech and in action with which an individual or a group is targeted. "Extreme acts such as mass murders or terrorist attacks are manifestations of hate," it adds [18].

When it comes to identifiers, most digital threats relevant to hate speech and online violence share common characteristics when they manifest online. These are some of the notable identifiers of hate speech:

1. The content targets a group, or an individual as a member of a group,
2. There is a dehumanizing sentiment within the content,
3. The content encourages and justifies a collective self-defence, which is also known as 'accusation in mirror'
4. There is a call to action in the content – murder, violence, segregation, discrimination, xenophobia and others.

Preparing your social media account

Having a rich and diverse timeline is the foundation to this research. Here are some of the main reasons in seeking a timeline that is inclusive of all the available voices on each platform:

- So that you can see events in their entirety (comprehensive coverage)
- So that you can find out what is trending easily
- So that you can form an objective opinion about the events,
- So that you don't miss what is happening,

It is important to note that people with unpopular views are your friends in this case, because they often tend to host content that is either harmful or can lead you to content that is harmful. This lays the foundation for monitoring trending keywords, which makes keyword generation a straightforward process.

For both Meta and Twitter, there are a few steps that call for a systemic approach in planning and preparing the account that would be dedicated for the task of monitoring and tracking contents. The first and most direct approach is to follow or 'like' pages, groups or profile accounts that represent diverse views and opinions using personal (own) social media accounts. Although this makes sure that all voices are represented in the timeline of your account, 'liking' or following such accounts pose significant challenges:

- You could be judged, put a label on, or ostracized for associating with such accounts,
- You will have too much content in your news feed overwhelming you and wasting your time,

A less conspicuous way could be to build a database of accounts along with the essential details on a spreadsheet and update the entries regularly. The challenges of using a database on a spreadsheet could be:

- It can build up easily and create additional work because you may need to update it continuously,

- The inconvenience of using a separate app for this

Another way of preparing and organizing these social media accounts is with *Custom lists*. Meta used to have features suitable for this task on its Facebook platform. However, it only limited the use of this feature to organizing friends and contents for specific people from friends lists. Users can also use this feature to manage privacy concerns by excluding some of their audience from their friends list from viewing individual content [17].

On Twitter, however, the *Lists* feature allows users to add accounts irrespective of their status. One can create up to 20 lists with a maximum of 500 names per list regardless of these names being friends or followers. The intention behind this feature is to arm a user with a capability to create distinct timelines based on interest, expertise, geography, or any another rationale [19]. Twitter's *Lists* feature is easy to find and use within the interface, which can be made private or public.

Identifying keywords

Once you have a rich and diverse timeline, it is time to search and find keywords. The foundation of threat tracking is built on keywords and being able to efficiently look for keywords is of paramount importance. Conducting keyword search enables a researcher to discover:

1. The identities of notable influencers,
2. The harmful sentiments associated with the keyword and the variations thereof,
3. The reach (spread) of the topic represented with the keyword.

There are various approaches to identify harmful keywords, but here are two approaches that can be used on SM platforms. For both approaches, the keywords to look for are names, acronyms, nicknames, special terms, and so on.

Approach 1: Using the *Comments* section to identify the exact keyword that holds the threat:

- Frequently monitor ALL the pages of accounts in question (including fake/impersonation pages),
- Take time to carefully go through the comments and you will find some keywords that you can search for and see the influence thereof. *NOTE: Try to avoid the knee-jerk reaction following your discovery of one BIG keyword. Sometimes, the first keyword you find is not necessarily the best one, or there is a possibility of capturing multiple keywords of similar importance from the same post. Hence, resist the urge and collect as much material as you can from the interaction.*
- Search the keywords one by one,
- Get the variations of these words (spelling and verb variations), especially if they are not in English,
- Make a judgement call and stop if your search returned empty or low-level reach. It is usually a waste of time.

Once you have decided to stop searching for the keyword, it is usually a good idea to write a short note about it. If the keyword reappeared later, you would be able to make an informed connection with the user or the targeted group. These keywords may contain: hate speech, bullying, harassment, violence, sexism, etc.

Approach 2: Identifying trends to come up with a keyword that could potentially take you to the harmful content:

A. Examine your timeline,
   - What are the trending topics?
   - What are influencers talking about?
   - What about the media, opposition groups, armed groups?
B. Monitor your timeline,
   - Analyze contents for the tone of exchanges for possible harmful sentiments
   - Identify and search the keywords
   - Use specialized tools like CrowdTangle (Meta, Instagram, Reddit and Twitter), TweetDeck, Personalized 'Trends for you' feature on Twitter, etc.

Because social media sites are an enormous jungle, it is advisable to join networks that have similar backgrounds. Joining fact-checking, harmful speech trackers is advisable not only because of the shared vision, but such networks are great knowledge hubs.

The duties of this activity can sometimes be a risky endeavor, which calls for online and offline privacy precautionary measures. The job of identifying and tracking means going after harmful speeches and not the speakers, but someone out there might think otherwise.

For multilingual countries, keyword searching on those communities you are monitoring can sometimes become challenging. However, an effective approach is to carefully consider the contextual, cultural, and linguistic nuances of keywords generated from a multilingual community during analysis.

Circumventing content moderation

Today, it is estimated that there are about 4.76 billion active social media users in the world – almost 60% of the total global population – exchanging billions on contents daily.

And growing number of social media users are now utilizing nefarious tactics to circumvent automated content moderation so that their content can remain online even if it is in violation of the platform's user policies. And while platforms' response to ensure compliance with their policies and community standards has arguably been reactive, the use of Artificial Intelligence against problematic contents is gaining momentum in influencing decision-making process of content moderations.

And learning the tactics and strategies that such social media users employ to evade moderation algorithms would help us identify the pitfalls that make identifying harmful contents more challenging. Some of these tactics and strategies include:

- Using ciphered language,
- Using keywords with lexical variations,
- Visual blocking of keywords in images (superimposing texts over images),
- The use of alphanumeric characters,
- Using screenshots instead of textual content or links,
- Putting additional information in comment section,
- Using the likeness of prominent personality and institution to spread harmful content,
- Camouflage malicious content within harmless looking links/ads (also known as "cloaking"),
- Using neutral or semi-neutral hashtags with borderline provocative tone,
- Switching to temporary or short contents such as Facebook/Instagram Story and unrecorded Twitter Space,
- Redirecting users using third party features such as Linktree and Bio,
- Briefly flashing harmful imagery within a longer video,
- Streaming harmful contents live,

It is worth noting that researchers engaged in monitoring and tracking harmful content can progressively learn to better identify such behaviours by upskilling themselves while using a combination of technical and social research skills**.** Additionally, some sophisticated attempts may require researchers to be vigilant and maintain high-level awareness about local and international affairs so that they can consistently and reliably identify trends that could trigger harmful content to come into being.

It is important to work collaboratively with partners, stakeholders and other researchers to develop effective strategies for monitoring and tracking harmful content. Below is a list of non-exhaustive points that researchers may consider as part of an effective monitoring and tracking strategy:

1. Having a comprehensive knowledge of user policies and community guidelines will help researchers identify behaviours that are prohibited and provide a framework for assessing content on respective platforms.

2. When possible, use third-party tools and technologies that can help identify patterns of behaviour that are suspicious or anomalous. For example, there are inexpensive monitoring tools that use effective machine learning algorithms to identify language patterns that may indicate a user is attempting to circumvent AI moderations.
3. While using automated third-party monitoring tools, researchers can put in place filters (also known as rules) that are programmed to send instant notifications when selected keywords appear on the platforms being monitored.
4. Monitor social media platforms for changes in behaviour, especially when during changes and high-profile events. For example, during an election cycle or a period of civil unrest, you may see an increase in users attempting to utilize tactics to circumvent content moderation.
5. Look for inconsistencies in content, such as sudden spikes in activity or unusual patterns of behaviour that may indicate coordinated or malicious activity.
6. Engage with the social media community to gather insights into how users may be attempting to circumvent content moderation. Interactions such as comments, shares and retweets could provide valuable insights into the strategies bad actors are employing.
7. Look for specific attributes of the content, which could be regarded as typical evasive behaviours :
    - Language: patterns of coded language or ambiguous references may be used to evade detection by algorithms or human moderators.
    - Context: examine the context in which content is posted, such as the time, location, or user demographics, to identify patterns that may indicate coordinated or malicious activity.
    - Amplification: a content could be amplified through fake accounts or third-party services to evade detection.
    - Image or video manipulation: Look for alterations or manipulations of images or videos that may be used to evade detection by algorithms that recognize explicit or hate speech content.
    - Evasive tactics: a content could use tactics like conceal offensive content in seemingly innocent posts or comments.
    - User behaviour: Monitor the behaviour of individual users to identify patterns such as cross-posting multiple contents rapidly, using multiple accounts, Keyword Stuffing (artificially boost the relevance of a piece of content for a particular keyword or phrase).

In summary, it is important to approach this task with a critical eye and remain objective throughout the process. It is also essential to prioritize user privacy and not engage in activities that may violate user rights or lead to unethical research practices. Ultimately, the goal of monitoring online content is to identify, track and document problematic content, while promoting a safe and positive online environment for all users.

Privacy and Safety of Researchers

Protecting the privacy and identity of individuals who monitor and track online contents is crucial to ensure their offline safety and well-being. Here are some steps that they can take to protect their privacy and identity:

1. Secure work environment: institutions and civil society organisations (CSOs) should provide their researchers a secure and confidential work environment that is free from any potential threats or harassment.
2. Use pseudonyms: it is important researchers are assigned pseudonyms instead of their real names, which can help protect their identity.
3. Restrict access to researcher information: access to the researchers' personal information like their name, addresses and other contact details should be restricted to only a small circle of individuals.
4. Use secure communication channels: everyone involved should use secure communication channels to communicate with each other as well as stakeholders, such as end-to-end encrypted messaging apps.
5. Provide mental health support: researchers are often exposed to disturbing and traumatic content, which can take a toll on their mental health. Mental health support should be available to help them cope with potential consequence of emotional and psychological distress.
6. Regularly update privacy policies: the privacy and safety policies should regularly be reviewed and updated to ensure that they remain effective in protecting the privacy and identity.

By adding these measures to the internal security and safety rules of organizations can help ensure that researchers are safe and protected, and that they can perform their important role without fear of harm or harassment.

Tracking Online Contents

With the rapid growth of information and communication technologies, social media tracking tools have become increasingly user-friendly. [7] Most of these trackers offer assorted data, comprehensive analysis, and visualization capabilities that provide context to the systematically collected information. [6] However, tracking social media posts and accounts should ideally begin with a preparation phase where partners would conduct quick research to identify keywords as well as the main actors to track. [3]

Once this is setup, it becomes less challenging to begin collecting data using various methods (see below). Even though the data type from each platform may differ, these APIs provide detailed information of a particular entry such as number of likes, shares, comments, reactions, views, retweets, and quotes. In some cases, information about other media forms, including links, images and videos is also retrieved along with the date and time of posting.

Data collection

Social media data collection can be accomplished within the premises of two major categories: direct and indirect access. [22] Collecting information directly from the web page's source code, commonly known as "web scraping", provides direct access to the available data. Accessing data via third-party tools is an indirect way of collecting data from social media platforms, while using an Application Programming Interface (API), which enables two software components to communicate with each other so that accessing certain aspects of a platform's data would be possible.

There are benefits and challenges to approaching data collection either directly or indirectly. According to a document published by National Democratic Institute, collecting data via respective platforms' APIs "enables a user to quickly process much more data, and in turn generate deeper insights into online activity, than they would otherwise be able to do manually."

Additionally, APIs are not designed for a particular programming language, which makes their applicability overarching. The most common languages one can use to retrieve API data are Python, R, and Java. However, the benefits of using APIs come with a limitation in the volume and rate of data collection, which most social media platforms insist is due to wanting to manage the large volume of these requests and system performance and stability. [20] Twitter, for example, has set a cap on the volume of collection depending on five access levels – from 'Essential Access' to 'Enterprise Access'. [30]

	Historical API available	Streaming API available	Third-party data collection tools
Twitter	Yes	Yes	CrowdTangle (extension)
Facebook	No	No	CrowdTangle
YouTube	Yes	No	NA

Steps to Collecting Data

R is a software environment and statistical programming language built for data analysis. Its capabilities can broadly be summarized into three categories: manipulating data, statistical analysis, and visualizing data. It is considered one of the easiest programming languages to learn, as anyone with little to no knowledge and experience can perform the basic tasks in the context of social media monitoring and data analysis.

Using R-studio to collect data from social media platforms is fundamentally similar. Hence, this document will provide a step-by-step guide on how to use this programming tool to collect data from the Twitter API using R-Studio.

Step 1: Download and install R and R-Studio for Desktop

Choose a version of R and R-Studio from their websites that matches your computer's operating system. The software includes a console, syntax-highlighting editor that supports direct code execution. It also features tools for plotting, viewing history, debugging and managing your workspace.

For superior results, having a 64-bit operating system is preferred. However, for 32-bit computer systems, older versions of R-Studio work just the same. Please consult the website as some computer systems may need installers and additional files.

Step 2: Getting and using a Twitter developer account

In order to collect data from Twitter, the first thing to do is to apply for a developer account. This is a secure means of using your account to interact with the platform through codes, which used to be free and no approval was needed. However, researchers need to apply for an account on its developer petition page (https://developer.twitter.com/en/portal/petition/essential/basic-info).

After creating an account on Twitter's website, researchers should create a Twitter API App. According to Twitter's user guideline (https://help.twitter.com/en/rules-and-policies/twitter-api), when someone wants to access Twitter's APIs, they must register an application. By default, applications can only access public information on Twitter; certain endpoints—such as those responsible for sending or receiving Direct Messages—require additional permissions from you before these APIs can access various data from the five generically categorized endpoints: Accounts and users, Tweets and replies, Direct Messages, Ads, and Publisher tools and Software Development Kits (SDKs). These APIs can retrieve or create a variety of different resources, including tweets, users, spaces, direct messages, lists, trends and many more resources.

The last step in this process is to start pulling tweets from the Twitter API using R-Studio. Here is a summary of the steps involved that is designed for a beginner in collecting data from social media using just a few lines of codes and commands:

- While signed on your Twitter account, go to https://developer.twitter.com/en and sign up,
- Agree to Twitter's term of signing up; its "Developer agreement & policy",
- Verify your Twitter Developer Account through the verification email sent,
- Back at the Developer Portal page, click "Create New App" and give a name to the new App,
- Give a unique name for the new App (it can be changed later),
- Three keys will be provided:
    1. **The API Key:** Think of the API key as the username that represents your App when making API requests. It helps us verify who you are.
    2. **The API Key Secret:** Your API Key Secret is like a password and helps verify your API Key. This will be one of the last times you'll see it displayed, so remember to save it in a safe place.
    3. **The Bearer Token:** it is An Access Token used in authentication that allows you to pull specific data.
        - These keys verify and allow you to make requests to the Twitter API.
        - For security, this will be the last time we'll fully display the Secret.
        - Save them in a secure location
        - Treat them like a password or a set of keys. If security has been compromised, regenerate them
        - DO NOT store them in public places or shared docs
- Go to the App settings button and setup the user authentication settings, which can be changed anytime. This authentication allows users to log in to your App with Twitter. It also allows your App to make specific requests for authenticated users. The following are found under these settings:
    - App permissions – enable permissions for read, write, direct message, or all,
    - Type of App – enables a web App, Automated App or Bot,
    - App information – specifies the destinations for the Callback/Redirect URLs following the authentication process,
    - Provide additional details such as website URL, organization name, organization URL, terms of service and privacy policy of the App

    - NOTE: Callback URL can be anything if you don't plan on using log in with Twitter to allow others to authenticate your app, even https://example.com will work. Website URL also doesn't affect anything else apart from being informative. It can also be anything. You probably don't need to set anything there because the read-only bearer token may be enough.
- A set of keys for 'Client ID' and 'Client Secret' will then be available, which should be saved in a secure location,
- On the next page, under the title of "Authentication Tokens," click the generate for "Access Token and Secret" and save the two codes in a safe and accessible location,
- Once you have retrieved your application, log in to Twitter on a browser and head to https://apps.twitter.com. You can retrieve your application's consumer and access keys here – there will be four keys. These are simply strings of text that your application will use to verify that it is indeed requesting data on behalf of a person – in this case, you. This usage of keys for authentication is a process known as oauth, which stands for open authorization. It developed as a way to allow applications to perform actions online on behalf of users without transferring their password and username for every action.
- Once you have your keys and your app name, you can use the code that follows in the screenshot below to create your authorization "token". Once this token is created, you are ready to start using the Twitter API.
- **NOTE:** Please see the detailed tutorial on how to get started with R on Twitter's website: https://developer.twitter.com/en/docs/tutorials/getting-started-with-r-and-v2-of-the-twitter-api

Data Analytics, Listening and Tracking Tools

The capabilities of automated content analysis have rapidly grown because of the technological advancements in machine learning and Artificial Intelligence. [12] This growth has also provided room for the creation of many social media analytics tools, many of whom have multiple use cases.

According to IBM, social media analytics is "the ability to gather and find meaning in data gathered from social channels to support business decisions — and measure the performance of actions based on those decisions through social media." [38]

It is not just quantitative analysis of measurements such as likes, follows, retweets, views, clicks, and impressions gathered from social media accounts. Datasets retrieved from platforms are sorted to extract "meaningful insights" that are expected to provide analysts with valuable information in their investigation of content and actors.

Once the desired keywords and actors are identified, analytics tools will collect, organize and analyze these entries, providing critical context within a specified duration. Some tools have advanced features that track and predict the future impact of content within the platform's community. Partners, who often have intimate knowledge of the offline context, would then be able to synthesize the interrelationships of online content with offline events.

Even though there are many data analytics tools on the market, advanced data analysis, like sentiment analysis, is in its early stage of development, mainly because of undeveloped capability with foreign languages. Nonetheless, below are some of the most prominent social media analytics, listening and tracking tools:

CrowdTangle

CrowdTangle tracks influential public accounts and groups across Facebook, Instagram, and Reddit, including all verified users, profiles, and accounts like politicians, journalists, media and publishers, celebrities, sports teams, public figures and more. CrowdTangle also can track seven days of public Twitter data via CrowdTangle Search and their Chrome Extension.

Acquired by Facebook – now Meta – in 2016, CrowdTangle provides detailed information about how content is being shared across Facebook, Twitter, Instagram and Reddit. According to CrowdTangle, here are examples of its use:

- Journalists using CrowdTangle Search to search across Facebook or Instagram for content relevant to their reporting.
- Social media managers tracking their own account performance and comparing themselves to the competition in Intelligence.
- TV producers broadcasting real-time streams of social posts related to breaking news events using Live Displays.
- Fact-checkers identifying posts that contain misinformation.
- Researchers analyzing trends across thousands of accounts over time and reporting on how information spreads.

Pricing: CrowdTangle is completely free, anyone can access its public features, the Link Checker Chrome Extension, as well as the Live Displays public hub. However, CrowdTangle only provides its full access to Facebook partners who publish original contents. This includes, according to the information on its website, "media publishers, music labels, public figures, and content creators, as well as the managers and agents who administer accounts on their behalf." Additionally, journalists, fact-checkers, academics and researchers also qualify for some access of its features under conditions that are specified on its website for each discipline.

Recommended Steps to using CrowdTangle

Step-1: Request for access

Users must validate that their account represents an authentic news source through Meta's vetting and review process. Below are some of the most important requirements for pages to qualify for Meta's news-page registration:

- Publish as a news entity, not an individual, and reflect the judgment of a multi-person editorial staff,
- Primarily create journalism that reports on current events or timely information, and that is not parody or satire,
- Be active in the last 90 days,
- Link to a functioning website with a verified domain,
- Confirm its ownership and identity through business verification,
- Comply with Meta's Community Standards, Advertising Policies and other relevant policies,
- Have the majority of its posted links come from the website verified for news page registration.

Additionally, fact-checkers qualify for CrowdTangle access if they are a part of Facebook's Third-Party Fact-Checking program, or if they are signatories of the International Fact-Checking Network (IFCN). Journalists or freelance contributors can also register as a journalist on Facebook using their personal account. And to be eligible for the registration, an applicant must live in one of the 38 countries where Meta's voluntary journalist registration is currently available. Some of the most notable requirements include:

- Applicant's Facebook profile has a profile image, cover image and at least one post,
- The name listed on the applicants Facebook profile is identical to, or closely matches, the name listed in the article bylines or on the publisher's staff directory,
- The applicant must be able to verify their affiliation with a registered news organization,

Generally, there are six steps to getting started with the registration process:

1. Add your website to your Facebook page,
2. Verify your business,
3. Verify your website's domain,
4. Submit your application,
5. Tell Meta more about your publication,
6. Grant Meta access to your paywall content,

Step-2: Create teams & dashboards

Both dashboards and teams can be managed from Account Settings that can be found under the gear icon on the upper right-hand side of the CrowdTangle dashboard. The 'Dashboard' section allows you to create new dashboards, see, and access all dashboards that have been created.

Teams are groups of people who need the same access to the same dashboards. This feature is one of the ways that can be used to organize the access to your Dashboard in your account.

Step 3: Set up local lists for pages and groups

Lists appear on the left-hand side of the dashboard. They are collections of specific pages, groups, accounts, or subreddits. Once a list has been set up, you can use it to surface content or perform analytics in CrowdTangle. Lists can be added to your dashboard in four ways:

1. Via the Explore section,
2. Creating a custom List,
3. Batch upload,
4. Importing Lists.

Monitoring events and accounts is about listening to who is adding to social conversations, and researchers use CrowdTangle to determine which accounts are driving more social conversations, identify new topics and influencers. CrowdTangle recommends the following ways for the respective activities:

A. Create a monitoring Live Display:
   - Decide which platforms are most representative of the actors/events in the event you're following,
   - For each platform you want to track, create a List for each actor (event) Pages/Profiles/Accounts,
   - In your monitoring Live Display, create a Leaderboard column for each of those Lists to allow you to see what is driving the most social conversation and what percentage of the overall interactions for the List comes from each candidate.

B. Identify the trend with Intelligence:
   - Create a List of the Pages/Profiles/Accounts of each actor/event in the election you're following.
   - Go to Intelligence in any Dashboard,
   - For each platform you want to track, add the Pages/Profiles/Accounts of each actor you're following for that particular event,
   - Use this to compare the performance of the different actors,
   - Export PDF data and share the visualized result.

C. Identifying new topics and influencers:
   - Create a 'Saved Search' in your Facebook and/or Instagram dashboards with your monitoring theme. For example, you can set up the search to look for names,
   - Create a Leaderboard column in your monitoring Live Display for each of the searches created,
   - Study the top entries in the Leaderboard and consider adding them to your existing Lists.

Once a List has been created, anyone with full Dashboard User privileges can edit it. To edit an existing List, simply click on the name of the List under the List section on the left side of your screen, and then click on "Manage".

Step 3: Setting up email notifications

Notifications are a way to have important information from a CrowdTangle Dashboard sent via e-mail or Slack to you or those around you. Notifications can be created, edited, or deleted in the Notification section of CrowdTangle, which can be found in two places: either in the Notifications tab of a List or Saved Search or in the global Notifications section on the left-hand navigation bar. Here are some of the reasons you would want to use Notifications:

- If you want to know what your top overperforming posts have been every week,
- If you want Leaderboard data sent to your team before monthly ops reviews,
- If you'd like to identify high performing content in real-time,

- If you want to be notified when an influencer is linking to your content

Please note that only full Dashboard users can create or modify Notifications, while users with limited access can only subscribe to existing Notifications.

Talkwalker

This is a Consumer Intelligence Acceleration platform that features AI-powered tracking tools, which focus on social listening. It tracks, identifies, and analyzes contents and trends. Some of its features include:

- AI-powered intelligent content and trend analysis,
- Social intelligence monitoring for customer and market insights
- Quick campaign and social channel management,
- Social Content Ratings (SCR) for social performance bulk measurement,
- Flexible Boolean searches with smart filters for data fragmentation

Pricing: Talkwalker's subscription plans are divided into four categories: Strategic, Premium, Business, and Starter. Though Talkwalker does not provide pricing for each of its plans, the basic plan is reported to cost between $9,000 and $9,600/year.

The user manual that Talkwalker developed for beginners describes the process of setting up a social media monitoring project, which is divided into five key steps. The manual advises its users to take great care while planning each step:

Step 1: Search

The starting point of any social media monitoring project is defining the keywords and particular social feeds you want to monitor. A wide range of search operators can be added to your search queries to ensure that you only get relevant results without wasting time and resources.

Therefore, here are some of the key topics to search for as a first step:

- **Names of actors and key people involved:** Make sure to search for all variations, common spelling errors and social media handles.
- **Keywords:** Monitor the social and online discussion around keywords linked to the topic you are monitoring to understand the key issues around it.

Take the time to write down your search terms and to fully test the accompanying search input with your social media monitoring tool. By using simple Boolean Operators like AND, OR and NEAR/X, you can refine your searches quickly and easily.

For example, using the following syntax in the search query "chocolate cake" AND cream NOT "ice cream" will optimize searches to return results of **chocolate cake** that only include references to **cream** and not **ice cream**.

Step 2: Monitor

With your search queries set, simple monitoring can begin, which will show you the volume of mentions and how this has fluctuated over time. With monitoring, you can identify and document trends over a certain period of time where outliers, in an otherwise consistent flow, are indicators of special events.

For Talkwalker, the key thing to remember at this stage is that this is purely a first look to get an idea of how your topics and keywords are performing. To get accurate results, we need to filter the results even further.

Step 3: Filter

Once your results are coming in, it's time to filter to find relevant insights. Results can be filtered by country, language, sentiment, Twitter follower numbers and more, in which customizable filters are also available.

To ensure quick and consistent qualification of results, it is important to determine the evaluation criteria beforehand. To allow for flexibility and extensive customizing options, researchers should use Talkwalker's filtering options:

- Language & Country/Region: Because it is rarely necessary to search the whole social web, you should only select the language(s), countries and regions that are relevant to your project.
- Media Types & Source Panels: Not every website, source or media type will be relevant. It is, therefore, advisable to exclude sources that are not relevant or of poor quality. Conversely, certain sources will be more important than others, which can be included in a pre-defined source panel.
- Influence & Relevance: The ways of determining influencer impact vary according to the platform being measured. Together with other factors, such as the frequency of interaction with the content of a source, relevancy can be customized to meet particular needs.
- Sentiment: The most commonly used qualitative factor is the sentiment or "mood" of an event, which is derived by analyzing the tone and attitude contained within the text of a content. Talkwalker uses a general classification of positive/neutral/negative.

Step 4: Analyze

With results filtered to include only the most relevant, analysis can begin. Look at demographic insights, compare results by media type, understand increases in mentions and see how certain posts went viral. Talkwalker's Analytics feature provides insights once your keywords are identified, basic monitoring began to take shape and filters have been implemented.

The Smart theme feature allows you to see, for example, public figures or general topics and themes linked to a topic, keyword or hashtag. Customized themes can help you to see how often keywords pivotal to your monitoring are being mentioned alongside other relevant parameters. Here are few of such features included in Talkwalker:

- Influencer Tables: Influencer tables help you to understand which websites, blogs, Twitter accounts, Facebook pages and more are the most influential commentators about a brand, product or industry.
- Demographics: Data such as the languages, countries and gender breakdown of social discussion helps to get a clearer idea of where social discussion is taking place. Such insights can help companies target campaigns and communications more effectively and understand where they have the most influence.
- Sentiment Breakdown: These are ideal for getting a quick idea of the potential scale of a crisis and can also be useful for measuring public reaction to a new campaign.
- Dashboards: The primary function of dashboards is to allow you to see the analytics that matter for you in a layout that makes sense. Having clear, intuitive data visualizations is a key element of good analysis. Both pre-set dashboards – like Talkwalker IQ Apps – and customizable dashboards help you to see all the metrics and analytics you need in one place.
- Virality Maps: These maps show where and when messages spread and which sources were most important in helping them spread. They help you understand why and how a particular post, tweet or article spread across social media and online.
- Spike Analysis: It involves focusing on particular spikes in mentions and then drilling down into the data to understand why such an event occurred. This type of analysis is particularly useful to understand the causes of a crisis, why certain campaigns were successful and to better understand the types of posts, stories and campaigns that capture the social public's attention.

Step 5: Report

Once insights have been uncovered, the final step is reporting. Having good reporting functions is critical as they not only allow you to distribute insights to all relevant stakeholders, but being able to create reports quickly also helps to save the time needed for analysis and reporting, according to Talkwalker's client tests.

All relevant charts/graphs can be selected for inclusion in a report and then scheduled on a daily, weekly or monthly basis. Creating alerts based on irregular activity also helps to warn you of an impending crisis or an unexpected positive event.

Keyhole

This is a real-time social media tracker tool that offers automated social media analytics for agencies, marketers, journalists, and executives. The goal is to help users accurately measure social data, both real-time and historical. In any case, the information is visualized and kept simple, in graphs, layouts, and numbers. The tracking process focuses on hashtags, keywords, URLs, and social media accounts on Facebook, YouTube, Instagram, and Twitter. Some of the features include:

- Influencer Tracking and Monitoring
- Social Media Account Monitoring and Reporting
- Publishing & Scheduling
- Hashtag and Keyword Tracking and Analytics
- Twitter Historical Data
- Social Media Analytics API

Pricing: Keyhole no longer provides pricing information for all its plans except the two most basic ones:

1. Profile Analytics – From $39/month
2. Social Listening & Campaigns – From $79/month
3. Influencer Tracking – custom pricing
4. Enterprise Suite – custom pricing

Below are the basic steps to begin working on Keyhole:

Step1: Sign up:
Enter the URL www.keyhole.com and click on the start free trial option or subscribe for one of Keyhole's plans.

Step 2: Enter personal details:
Enter the first name, last name, email ID, password and contact number; select your industry and accept the terms and conditions then click on the Let's go button.

Step 3: Select your primary objective:
Select one of the options that you plan to use Keyhole for: monitoring, tracking, researching and more.

Step 4: Refine your primary objective:
Here we can get three choices. First, track mentions of a brand or product; secondly, track performance of a social media account; and lastly, track mentions of a market leader or industry topic.

Step 5: Create trackers and sub-trackers:
Here you can select Twitter, Instagram, Facebook, or YouTube and enter the account details. Trackers allow you to monitor an account in order to discover its performance for the last year. You can monitor the account, create instant reports to compare accounts, and understand how contents have become more engaging. Organize your trackers data into sub-trackers for easier analysis of individual keywords.

Step 6: Analyze and report:
This is the last step where you would analyze the monitoring data of the account and report. Keyhole will provide the most important analytics metrics and, depending on the status of your account, you will have capabilities to analyze the quantitative and qualitative measurements of the account you are monitoring, including impression and engagement rates.

NewsWhip

NewsWhip is a real-time media monitoring platform, which combines real-time feeds of web and social content with public engagement data to identify and predict the content that matters. The platform uses advanced algorithms and data analytics techniques to track the performance of online content across a variety of outlets, including social media, websites and blogs.

It can be used for media monitoring, crisis management, as well as research and strategy. According to the platform, its users range from journalists, content strategists, marketing executives, social media managers, and so on, across various industries. In general, to start using the NewsWhip platform, you are required to purchase membership to create an account and connect your social media accounts or other sources of media content. Once you have done that, you can access the tool's features and metrics to track and analyze the performance of your content. This may involve using the tool's dashboard, setting up alerts and notifications, or creating custom reports and visualizations.

Pricing: There is very little information about the pricing plan and model that is offered by NewsWhip, and it encourages its potential customers to contact the platform and request a demo on its website so that they can access its current pricing model and the associated cost. However, an unverified report suggests that its pricing starts from $2,000 per month per user.

The following beginners' guide can be used as a starting point:

1. Subscribe and log in to NewsWhip website,
2. Create dashboards that will enable you to view multiple widgets for one or more searches simultaneously. Widgets are the constitutional parts of a dashboard that let you choose how to view and display the results of your searches. Widgets can be mixed and matched across searches and dashboards to suit your needs. There are three main types of widgets that this platform employs:

   Feeds widgets: They present a feed of posts from the different platforms we track. This includes all web articles and social media, including Facebook, Twitter, Instagram, Reddit, and YouTube.

   Leaderboard widgets: These widgets gather data from across many posts and present that on an aggregate level. They display what websites, authors, or communities are generating the most interest for your search.

   Timeline widget: This widget shows how the media and public have engaged with a given topic over a specified time period. This powerful widget can be used to understand how interest in a crisis, event, or topic is developing and evolving over time.

3. The easiest way to get started is to use a preconfigured template with one of its three curated searches. The platform offers three separate templates that can also be selected for a custom search or with one of the platform's curated searches, which can give you access to a library of pre-built searches, ranging from key events, to topics, to trends that may be of interest to you. The three templates that can be optimized for their respective use case are:

   Monitoring Template: This template is ideal for monitoring a topic across the web and all major social networks.

   Crisis Template: This can be used to track how media and public interest in a crisis is evolving.

   Research Template: This is best used to understand how interest in an event or topic is changing over time.

4. You can expand each widget to display and look into the information contained in better detail. While you can drag-and-drop widgets in your preferred order, the platform provides a number of sorting and filtering options, among which is sorting by a time range where it can be configured by custom date range.
5. After refining the topics to search, begin working on the filters, including sources of information, language, locations and more.
6. After adding your search and refining it further to meet your needs, follow the basic rules of building a search to get the most accurate data as quickly as possible. This involves creating targeted search queries that can be done in one of the three ways: by keywords, by sources or by using URL links. It is important to understand the power of using the Boolean Search technique, which connects individual keywords or phrases within a single query in a number of ways.
7. Once you start the search, you will see results all categorized according to your inputs and filters, which is likely to indicate the types of additional filters that are needed.

8. By using the Metrics feature, you can measure the number of social interactions that each web article and post generates in a given time. The platform provides few metrics that generally analyze interactions, spread speed and performance.
9. The NewsWhip Spike dashboard also provides predicted interactions metrics by measuring social interactions on each content at regular intervals to create a time-series of interaction values. It will then predict interactions at a given time in the future after applying its predictive analytics technique on these values.
10. The platform offers the capability of sharing searches and dashboards with members of your team to work collaboratively.
11. Depending on the type of widget, you can export data from respective widget into PowerPoint, Excel, image, PDF, or JSON,
12. Additionally, the platform also offers the capability of setting up real-time alerts as well as scheduled alerts. Real-time alerts allow you to react faster with new articles and social media posts being discovered in as little as 90 seconds. It is a type of alert that is triggered by variables that you set beforehand. Scheduled alerts help you stay informed, giving you and your team a digest of what's new from as often as every 30 minutes.

Tableau

Monitoring harmful contents posted on social media can be done with Tableau, an intelligence and data visualization application. Utilizing its data analytics and visualization tools, it is possible to examine social media data and spot potentially dangerous information. The text of messages can be examined using its text analysis and sentiment analysis functions; for instance, to find posts with potentially objectionable or dangerous content. With the help of Tableau, you can make a range of dynamic and aesthetically pleasing data graphs and dashboards.

Additionally, Tableau's geospatial analysis tools can be used to spot regional patterns and trends in social media data, which can be utilized to pinpoint regions with a higher concentration of dangerous content. Tableau's dashboarding and reporting features can be used to build interactive visualizations and reports that assist in monitoring and tracking the spread of potentially hazardous content on social media once the possible harmful content has been recognized.

Pricing: Tableau’s subscription plans are divided into three categories: Tableau Creator (USD70/mo.), Tableau Explorer (USD42/mo.), and Tableau Viewer (USD15/mo.).

Here is a step-by-step guide on how Tableau can be used to monitor harmful content on social media:

1. On Tableau’s website, subscribe and create an account,
2. Once logged in, go to the "Connect" tab and select the data source that you want to track. There are two ways to do this: use the built-in connectors for popular social media platforms such as Twitter and Facebook, or import data from a CSV or similar data file.
3. Go to the "Data" tab and drag and drop the relevant columns into the workspace, which creates a visual representation of the connected data, such as a bar chart or scatter plot.
4. The next step is to clean and organize the data for analysis. This stage includes the following tasks:
    - Removing data that are irrelevant and unnecessary,
    - Merging data that came from various sources, and
    - Creating new parameters and measurements.
5. By selecting the "Filters" tab, refine the data by selecting the relevant criteria that would potentially return the specified contents, actors, keywords, phrases, social media platforms, languages, or locations.
6. In the “Analytics" section, go to the "Create Calculated Field", which enables you to track the performance of the selected filters, such as the number of likes, comments or shares.
7. The dashboard, while using it to monitor the contents and actors, can be configured to send notifications when certain conditions are met to your preferred address. These configurations can be set in a way that alerts are sent to you when, for example, a specified number of harmful posts are identified in a specific geographic area. This ideal feature helps in the automation of tracking accounts and users even outside of your normal working hours, which can help you to respond quickly to potential crisis situations.

8. This tool also possesses sentiment analysis capabilities to identify potentially harmful contents in the collected data. This is a built-in feature that uses word clouds and key phrase extraction techniques to identify common terms and phrases that may hold negative or offensive sentiment.
9. By using its geospatial analytics feature, you can identify the geographical patterns and trends found within the collected data. This feature can also create maps and other visualization means that show the geographic distribution of your data, as well as any other patterns that may be relevant to your analysis.
10. The capabilities to create interactive visualization reports will help you monitor the spread of contents over time.
11. Tableau also has collaboration features where you can share data about your findings with others – team members, partners and platform executives – and help them understand the context and prevalence of the topic being monitored.

Brandwatch

Brandwatch is a social media monitoring and analytics platform that can help researchers to monitor, track and analyse social media content. The platform uses algorithms and natural language processing (NLP) to collect data from various online sources. This data is then organized and analyzed to produce analysis and insights.

One of the key benefits of using Brandwatch is that it allows its users to track and analyze large volumes of data in real-time. This is very helpful as it enables them to understand the depth and spread of sentiments behind a harmful content, which in turn allows for quick response time. Additionally, Brandwatch provides visualization tools, including customizable dashboards, sentiment analysis, and real-time alerts:

1. Customizable dashboards: You can create personalized dashboards that display the most relevant data and metrics, which makes it easy to monitor and track contents using distinct indicators and their changes in time.
2. Sentiment analysis: The platform uses algorithms and NLP to analyse sentiments to categorize them as positive, negative, or neutral. This allows researchers to quickly identify trends and understand how people are reacting to the sentiments.
3. Notifications and Alerts: The platform can be configured to send notifications and alerts when specific keywords or phrases are mentioned online.

Pricing: The full product suite of Brandwatch has four categories: Consumer Intelligence for researchers and analysts, Social Media Management for social media managers, Influencer Marketing for influencer marketers and Essentials for small businesses. The essentials package, which it recommends is suitable for one to two users, is priced at $108 per month. However, it has not provided pricing information for its other packages that appear to have a quotation-based pricing model.

In general, the steps for using a platform like Brandwatch to monitor and track harmful online content might include the following:

1. Sign up for a Brandwatch account and log in to the platform.
2. Identify the specific type of content you seek to monitor by specifying it, whether it is hateful content, cyberbullying, or harassment.
3. Next, define the criteria for your monitoring activity on the dashboard. Use keywords, phrases, actors, or similar parameters to define and set up the criteria. Try using specific words and their common misspellings associated with the type of content you seek to monitor, which should also include their sources (such as social media platforms, websites, blogs…).
4. Use the platform's keyword filters to set up alerts to help you quickly identify and flag any potential harmful contents.
5. Monitor these alerts and review the flagged content to make a determination. Identify relevant contents as well as actors that match your monitoring criteria.
6. By using its advanced search tools and filters, narrow down the results further. It may sometimes be necessary to manually review the result to tweak the filters.

7. The platform describes four steps to producing actionable insights during this activity:

   Search: By using its versatile and AI-powered search to instantly find relevant data on any brand or topic, which gives you access to historical and real-time data from its fast-growing dataset.

   Segment: The platform gives you the capability to categorize searches by feedback, complaints, opinions and more filters through its machine learning classifiers that automatically segment the collected data.

   Analyse: The platform uses an AI analyst called Iris, which is optimised for analysing fast-changing and unstructured data, providing insights quickly. Its image analysis uncovers objects, scenes, actions, and logos in every image.

   Act: You can create, export and share data-driven reports in Excel, PPT, PDF, or via the Brandwatch API. You can also send live data to any screen that members of your team are working on. If you decide to publish a report, the platform enables you to measure how many people view, engage with, and rate your reports.

Advocacy

Monitoring and tacking of harmful contents or actors cannot be an end by itself. Social media researchers and advocacy groups, for example, can use the data that these tools generate to create effective community advocacy campaigns by employing targeted messaging coupled with outreaching efforts in raising awareness about the offline harm of the online contents.

On the other hand, however, the verifiable datasets and the subsequent contextual analysis present an opportunity for researchers to advocate for standards and policy changes within the social media platforms. And because the produced information is expected to be credible that is backed by data, positively engaging with social media platforms towards better and relevant policies could be possible.

There are several practical measures that advocacy groups can take to launch an effective campaign using the contextually accurate data analytics produced with the help of these social media monitoring and tracking tools. The measures, among others, should include:

1. Clearly defining the specific objectives of the campaign,
2. Developing a clear message that addresses the concerns of the victims,
3. Preparing a detailed plan of action on how to use the data analytics to garner social backing for the campaign,
4. Engaging directly with social media platforms by providing evidences of flaws in their policy and practices,
5. Being thoroughly prepared to respond to challenges from various actors, including social media users during the campaign,
6. Maintaining a constructive and transparent engagement with social media users, influencers and other stakeholders about the process that is guiding the campaign efforts,
7. Publishing reports, educational and awareness raising resources based on the initially set goals and objectives,
8. Using the same tools to monitor the effectiveness of the campaign and make necessary strategical adjustments.

Useful Resources

Internet Lab - https://internetlab.org.br/en/

Internet Lab is an independent research center that aims to foster academic debate around issues involving law and technology, especially internet policy. Our goal is to conduct interdisciplinary impactful research and promote dialogue among academics, professionals and policymakers. We follow an entrepreneurial non-profit model, which embraces our pursuit of producing scholarly research in the manner and spirit of an academic think tank. As a nexus of expertise in technology, public policy and social sciences, our research agenda covers a wide range of topics, including privacy, freedom of speech, gender and technology.

Dangerous Speech Project - https://dangerousspeech.org/

The Dangerous Speech Project was founded in 2010 to study speech (any form of human expression) that inspires violence between groups of people – and to find ways to mitigate this while protecting freedom of expression. The project works primarily in four areas: (1) Tracking and studying dangerous speech in many countries; (2) Researching effective responses to dangerous speech and other forms of harmful expression; (3) Advising social media and other tech companies on their policies, and encouraging them to engage in transparent research; (4) Teaching dangerous speech ideas to a variety of people who use them to study and counter dangerous speech.

Floating Sheep – http://www.floatingsheep.org/

Is a site dedicated to mapping and analyzing user generated geocoded data. The results provide one glimpse of what internet users (in the aggregate) think about particular places.

Demos – https://demos.co.uk/research-area/casm/

Demos is dedicated in digital research hub that provide unique insights and expertise across tech policy and its impact on our society, economy and democracy. CASM, in a joint venture with the University of Sussex, has pioneered the use of machine learning to interpret 'natural language' big data. This enables Demos to understand people, and conduct social research, at an unparalleled scale.

Tech Without Violence - https://techwithoutviolence.ca/

Developed by the Purple Sisters—a group of Ottawa youth dedicated to bringing attention to issues facing young women in their community—this new online hub provides practical information for young social media users and key recommendations social media platforms can use to address and prevent cyberviolence. This project focuses on five of the most popular social media platforms: Facebook, Twitter, Snapchat, Instagram and Tumblr.

Armed Conflict Location & Event Data Project (ACLED) - https://acleddata.com/

The Armed Conflict Location & Event Data Project (ACLED) is a disaggregated data collection, analysis, and crisis mapping project. ACLED collects information on the dates, actors, locations, fatalities, and types of all reported political violence and protest events around the world. The ACLED team conducts analysis to describe, explore, and test conflict scenarios, and makes both data and analysis open for free use by the public.

References

1 - "Bodyright - Own Your Body Online | Bodily Integrity | UNFPA." United Nations Population Fund. N.p., n.d. Web. 20 Oct. 2022. <https://www.unfpa.org/bodyright/>.

2 - Bradshaw, S., and P. Howard. "Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation." Computational Propaganda Research Project, Oxford Internet Institute, 2017, pp. 1–37.

3 - Brady, Madeline. Lessons Learned: Social Media Monitoring during Elections Case Studies from five EU Elections 2019-2020. June 2020. <https://democracy-reporting.org/en/office/EU/publications/lessons-learned-with-social-media-monitoring>.

4 - "Cyberviolence." Cyberviolence. N.p., n.d. Web. 20 Oct. 2022. <https://www.coe.int/en/web/cyberviolence/home>.

5 - "Cyberviolence | Tech Without Violence." N.p., n.d. Web. 23 Oct. 2022. <https://techwithoutviolence.ca/cyberviolence>.

6 - D'Ignazio, Catherine. "Data Visualization." The International Encyclopedia of Media Literacy. Ed. Renee Hobbs and Paul Mihailidis. 1st ed. Wiley, 2019. 1–10. DOI.org (Crossref). Web. 10 Nov. 2022. <https://onlinelibrary.wiley.com/doi/10.1002/9781118978238.ieml0095>.

7 - Fensel, Dieter and Universität Innsbruck, eds. Common Value Management: 1st International Workshop on Common Value Management CVM 2012 at the Extended Semantic Web Conference 2012 (ESWC2012), May 27-31, 2012 Heraklion, Greece. Stuttgart: Fraunhofer-Verl, 2012. Print.

8 - Gagliardone, Iginio et al. Countering Online Hate Speech. Paris: United Nations Educational, Scientific and Cultural Organization, 2015. Print.

9 - Gillespie, Tarleton. Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press, 2019. DOI.org (Crossref), https://doi.org/10.12987/9780300235029.

10 - "Global Consortium Launches Three-Year Effort to Strengthen Internet Freedom in 50 Countries." Information Saves Lives | Internews. N.p., 10 Sept. 2020. Web. 20 Oct. 2022. <https://internews.org/global-consortium-launches-three-year-effort-strengthen-internet-freedom-50-countries/>.

11 - Hargreaves, Eduardo et al. "Biases in the Facebook News Feed: A Case Study on the Italian Elections." 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Barcelona: IEEE, 2018. 806–812. DOI.org (Crossref). Web. 20 Oct. 2022. <https://ieeexplore.ieee.org/document/8508659/>.

12 - Hase, Valerie. "Automated Content Analysis." Standardisierte Inhaltsanalyse in der Kommunikationswissenschaft – Standardized Content Analysis in Communication Research. Ed. Franziska Oehmer-Pedrazzi et al. Wiesbaden: Springer Fachmedien Wiesbaden, 2023. 23–36. DOI.org (Crossref). Web. 10 Nov. 2022. <https://link.springer.com/10.1007/978-3-658-36179-2_3>.

13 - "Hate Speech | Transparency Center." N.p., n.d. Web. 20 Oct. 2022. <https://transparency.fb.com/policies/community-standards/hate-speech/>.

14 - "Hate Speech Policy - YouTube Help." N.p., n.d. Web. 20 Oct. 2022. <https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436>.

15 - “How Countries Are Regulating Internet Content.” N.p., 3 Jan. 2016. Web. 20 Oct. 2022. <https://web.archive.org/web/20160103124414/https://www.isoc.org/inet97/proceedings/B1/B1_3.HTM.

16 - “How We Respond to Inauthentic Behaviour on Our Platforms: Policy Update.” Meta 21 Oct. 2019. Web. 10 Nov. 2022. <https://about.fb.com/news/2019/10/inauthentic-behaviour-policy-update/>.

17 - Johnson, Maritza, Serge Egelman, and Steven M. Bellovin. “Facebook and Privacy: It’s Complicated.” Proceedings of the Eighth Symposium on Usable Privacy and Security - SOUPS ’12. Washington, D.C.: ACM Press, 2012. 1. DOI.org (Crossref). Web. 24 Oct. 2022. <http://dl.acm.org/citation.cfm?doid=2335356.2335369>.

18 - Keipi, Teo, et al. “Online Hate and Harmful Content: Cross-National Perspectives.” Taylor & Francis, 2016. library.oapen.org, https://library.oapen.org/handle/20.500.12657/22350

19 - Kim, Dongwoo, et al. “Analysis of Twitter Lists as a Potential Source for Discovering Latent Characteristics of Users.” 2022 https://www.researchgate.net/publication/242465472

20 - “Limits & Best Practices - Marketing API - Documentation.” Meta for Developers. N.p., n.d. Web. 10 Nov. 2022. <https://developers.facebook.com/docs/marketing-api/insights/best-practices/>.

21 - Michele Mazza, Marco Avvenuti, Stefano Cresci, and Maurizio Tesconi. Investigating the difference between trolls, social bots, and humans on twitter. Computer Communications, 2022. : 23–36. DOI.org (Crossref). Web. 10 Nov. 2022. <https://linkinghub.elsevier.com/retrieve/pii/S0140366422003711>.

22 - Monaco, Nick and Daniel Arnaudo. "Data Analytics for Social Media Monitoring." National Democratic Institute for International Affairs, 2020. <https://www.ndi.org/publications/data-analytics-social-media-monitoring>.

23 - NATO. “DEEP ADL Portal Infographic: Troll Factories (Media - (Dis)Information - Security).” NATO. N.p., n.d. Web. 10 Nov. 2022. <https://www.nato.int/cps/en/natohq/topics_175693.htm>.

24 - Nations, United. “What Is Hate Speech?” United Nations. N.p., n.d. Web. 20 Oct. 2022. <https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech>.

25 - “OHCHR | OHCHR and Freedom of Expression vs Incitement to Hatred: The Rabat Plan of Action.” OHCHR. N.p., n.d. Web. 20 Oct. 2022. <https://www.ohchr.org/en/freedom-of-expression>.

26 - “Our Policy on Glorification of Violence | Twitter Help.” N.p., n.d. Web. 20 Oct. 2022. <https://help.twitter.com/en/rules-and-policies/glorification-of-violence>.

27 - Paz, M. A., Montero-Díaz, J., & Moreno-Delgado, A. (2020). Hate Speech: A Systematized Review. SAGE Open, 10(4). https://doi.org/10.1177/2158244020973022.

28 - Soundararajan, T. Facebook India: Towards the Tipping Point of Violence: Caste and Religious Hate Speech. USA: Equality Labs, 2019. Print.

29 - “The Global State of Digital in July 2022.” DataReportal – Global Digital Insights, https://datareportal.com/reports/digital-2022-july-global-statshot. Accessed 23 Oct. 2022.

30 - “Twitter API v2 Tweet Caps.” N.p., n.d. Web. 20 Oct. 2022. <https://developer.twitter.com/en/docs/twitter-api/tweet-caps>.

31 - “Twitter’s Policy on Hateful Conduct | Twitter Help.” N.p., n.d. Web. 20 Oct. 2022. <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>.

32 - “Twitter’s Violent Threats Policy | Twitter Help.” N.p., n.d. Web. 20 Oct. 2022. <https://help.twitter.com/en/rules-and-policies/violent-threats-glorification>.

33 - “Umati: Monitoring Online Dangerous Speech | Dangerous Speech Project.” Dangerous Speech Project |. N.p., 2 Jan. 2013. Web. 20 Oct. 2022. <https://dangerousspeech.org/umati-monitoring-online-dangerous-speech/>.

34 - “UNESCO Launches Countering Online Hate Speech Publication | UNESCO.” N.p., n.d. Web. 23 Oct. 2022. <https://www.unesco.org/en/articles/unesco-launches-countering-online-hate-speech-publication>.

35 - Vergani, M., Perry, B., Freilich, J., Chermak, S., & Scrivens, R. (2022). PROTOCOL: Mapping the scientific knowledge and approaches to defining and measuring hate crime, hate speech, and hate incidents. Campbell Systematic Reviews, 18, e1228. https://doi.org/10.1002/cl2.1228

36 - “Violence and Incitement | Transparency Center.” N.p., n.d. Web. 20 Oct. 2022. <https://transparency.fb.com/policies/community-standards/violence-incitement/>.

37 - “Violent or Graphic Content Policies - YouTube Help.” N.p., n.d. Web. 24 Oct. 2022. <https://support.google.com/youtube/answer/2802008/>.

38 - “What Is Social Media Analytics? | IBM.” N.p., n.d. Web. 19 Oct. 2022. <https://www.ibm.com/topics/social-media-analytics>.