Social Media Data in Conflict Research

PLEASE DO NOT CIRCULATE

May 2022

Abstract

What is the role of social media in conflict-affected zones? How are people getting data to study social media in the most difficult circumstances, and what are the unique challenges that arise in such environments? In this report, we aim to shed light on these questions. We provide an overview of the ways in which social media data is currently used in conflict research, as well as the main limitations facing researchers collecting and interpreting social media data within conflict-affected contexts. Finally, we provide suggestions around how data access can be expanded to better support conflict researchers in their vital efforts.

Executive Summary

1. Social media serves an important role across a wide range of conflict-affected areas, from those facing intercommunal tensions to those experiencing state repression, interstate or international intervention, or armed group violence. In some cases, social media is essential to tracking conflict outbreak and human rights violations; in other, it is highly influential in shaping levels of group polarization.

   Conflict researchers study social media either as a focus of substantive interest (i.e. how social media affects conflict or how conflict affects social media) or use it as a tool for collecting data and recruiting research subjects. Despite the wide range of data collection methods used – from social media APIs and scraping to surveys and experimental designs – researchers are still limited by the lack of access to data that social media companies collect. This includes richer data on impressions, engagements and recommendations, but also more precise information about the number and qualifications of moderators employed for each region or the performance of algorithms involved in the moderation process across languages.

2. The plethora of challenges that researchers face ranges from difficulties accessing or interpreting social media data to vital issues related to user security and data privacy. Not all of these challenges are unique to the study of conflict, but they are even more consequential for research conducted in the most adverse circumstances and with the most vulnerable populations.
   2.1. In terms of data interpretation, particular caution is required around issues related to coverage and non-representativeness of user demographics, unbalanced and lagged reporting, bias in moderation, non-human activity and manipulations, user-driven changes to available information, and use of encrypted messaging apps.
   2.2. In terms of data access limitations, we highlight issues surrounding location metadata, moderated or taken-down content, content reach and recommendations, as well as broader issues related to data accessibility.

3. Meeting the needs of researchers working in conflict-affected areas requires balancing expanding data access with preserving user security. How to do so effectively and carefully is a great challenge facing social media companies and the academic community; we contribute to this discussion with a set of recommendations outlined at the end of the report.
   3.1. Striking this balance may include providing access to social media data at different levels of granularity and aggregation, particularly when it comes to location information and moderated or taken-down content.
   3.2. There is a high need for building more user-friendly tools that allow downloading or visualizing data in contexts where technical infrastructure may not support the extraction and the analysis of large amounts of raw data.

3.3. Providing access to data on exposure, engagement and algorithmic recommendations, expanding data features that can extracted through the API, and sharing transparent information about the moderation process would enrich the types of questions and insights that researchers could provide.

Table of Contents

Introduction

Social media is only gaining in importance for academic research, both as a medium for research and an object of inquiry. Providing an opportunity for individuals to transcend geographical boundaries and consume or create content at an unprecedented pace, social media has become an important force shaping societal dynamics and political outcomes. Researchers try to capture this by extracting and analyzing data provided by social media companies, or by developing research designs that allow them to test the questions of their interest. As the influence of social media grows, so does the need for greater data access and rigorous independent research. Perhaps nowhere is this more important than in conflict-affected areas, where social media data can be used to track violence or inform efforts to prevent it. Indeed, social media can be a uniquely valuable tool for tracking conflict developments or gathering digital evidence of war crimes. Recent years, however, abound with examples of social media facilitating the spread of hate speech and violent incitement, with often lethal consequences.

This report provides an overview of the conflicts where social media does or could play a role; existing data and its limitations in conflict-affected environments; existing research that employs social media data; and suggestions for expanding data access. We first establish a typology of conflict situations covered, which include situations of intercommunal tensions, state repression, interstate or international interventions, and armed group violence. Given their influence and user base size, our discussion centers around activity on Facebook, Twitter, Instagram, Reddit, TikTok, YouTube and WhatsApp. These platforms differ in the level and the type of data access they provide, which we summarize and use to identify some of the main challenges related to gathering, interpreting and sharing social media data. Identifying challenges is only the first step to overcoming them. With that goal, we provide a set of specific recommendations around data access that would allow researchers to more effectively investigate some of the most complex questions related to conflict-affected areas.

Methodology

In our discussion of sources, we cover the most widely used platforms globally: Facebook, Twitter, Instagram, Reddit, TikTok, YouTube, and WhatsApp. We do not include local platforms beyond touching on research that exploits them, since they are generally either small or – as in the case of China's Weibo or Russia's VK – provide more limited access to researchers. We included in our review published articles and working papers that we determined are relevant to the topic of social media and conflict; draw on social media data, analyzed through quantitative methods; and focus on conflict-affected or post-conflict countries. Additional information about the process can be found in Appendix A.2. In developing our typology of conflicts and recommendations, we additionally rely on policy reports and writing, as well as past experience working with social media data.

1 Applicable Conflicts

Social media plays a role across a wide range of conflicts. The single most important determining factor of its influence is internet penetration: where social media is not widely available or widely used, it will not play a significant role in fomenting violence or polarization, and it will be of limited use in affecting or tracking conflict dynamics. This applies to a number of countries in conflict, including South Sudan and Yemen. In even more cases, rural-urban divides in internet access affect the role of social media subnationally. During the Russian invasion of Ukraine, for example, rural areas faced particular challenges in staying online – and thus in reporting their stories (Cheney, 2022). For many fragile and conflict-affected states, however, Facebook is "synonymous with the internet" (Guo and O'Neill, 2021).

We identified four types of conflict where social media can be particularly important: intercommunal tensions, state repression, interstate or international intervention, and armed group violence. These are not mutually exclusive. The Syrian Civil War, for example, involves intercommunal tension between Sunnis and Alawites; state repression, in al-Assad's brutality towards rebels and civilians; international influence, through the various foreign actors supporting combatants; and armed group violence, in the involvement of terrorist organizations like the Islamic State (IS). Levels of violence also vary within conflict type, from post-conflict states like Bosnia and Herzegovina to active wars, like Ethiopia's Tigray conflict. Distinguishing between these four dimensions, however, is helpful in understanding the varied role that social media plays in fomenting violence and polarization – and where it can be used to better track or respond to conflict.

1.1 Intercommunal tensions

Societies divided by religious, ethnic, political, or social tensions are some of the most vulnerable to the polarizing effects of social media. Harmful speech – including hate speech, incitement, and disinformation – can intensify existing divides, in some cases leading to offline violence. These messages frequently initiate with leaders. Most infamously, Facebook posts by Myanmar's military (the Tatmadaw) directly contributed to the Rohingya genocide, by exploiting preexisting tensions between the Muslim minority and the majority Buddhist population. These campaigns spread stories of alleged atrocities to foment violence: one involved spreading rumors of an imminent jihadist attack among Buddhist networks, while simultaneously disseminating false accounts of anti-Muslim protests among the Rohingya (Stevenson, 2018; Mozur, 2018). In Ethiopia, social media posts by President Abiy Ahmed have stoked tensions between his ethnic group, the Amhara, and ethnic Tigrayans (*DW*, 2021). During the May 2021 Israel-Palestine escalation, a spokesman for then-Prime Minister

Benjamin tweeted footage that he alleged showed Palestinian militants launching rockets against Israelis; in fact, the footage was from Syria or Libya (Frenkel, 2021).

At the same time, social media can be essential to tracking, or even reducing, intercommunal conflict. In regions with high levels of violence, social media may be the only method of reporting human rights violations and conflict outbreak. The Syria Justice and Accountability Center, for example, has used social media posts to build a dataset of human rights violations committed during the Syrian Civil War.[1] In the future, social media may also be used to predict conflict outbreak (Bazzi et al., 2019). As discussed later in this report, social media can also help bridge group divides between geographically segregated areas (Asimovic, 2021). However, it will be more effective in doing so within lower-intensity conflicts and post-conflict societies than in active war zones.

1.2 State repression

In conflicts that pit the state against civilians, social media is often crucial for coordinating opposition to repressive regimes. The Arab Spring, the 2019 Venezuela uprising attempt, and the 2020 Belarusian protests all relied on social media for sharing information. Recent work, discussed in greater detail later in this report, has used social media to explain and track protest movements (e.g. Won et al., 2017). Where media is state-controlled, platforms allow opposition to undermine regime narratives or to publicize human rights abuses by the state.

Repressive regimes, meanwhile, use social media to limit dissent or justify offline violence. Bots and trolls are frequently used to spread propaganda, which can both bolster support and make opposition appear isolated. For example, Venezuelan president Nicolás Maduro's government got pro-regime hashtags to trend through both inauthentic accounts and by paying regular users for retweets (Suárez and Ponce de Leon, 2021). In other cases, social media may provide a platform for governments to target or justify repression. Rodrigo Duterte's government in the Philippines has targeted opponents with harassment campaigns, often for speaking out against his brutal "war on drugs." For example, senator Leila de Lima, who was investigating the government's use of extrajudicial killings, became the target of a wave of disinformation: posts circulated claiming she took money from drug lords, and that doctored her face into pornography (Alba, 2018). The Duterte government also uses social media for "red-tagging" opponents as communist rebels, which often leads to violence against those named (Human Rights Watch, 2022).

[1]See https://syriaaccountability.org/database/ and https://syrianarchive.org/en/about.

States additionally use offline tools to reduce online dissent. In countries like authoritarian and hybrid regimes, including Nicaragua, Saudi Arabia, Hungary, and the Philippines, laws aimed at criminalizing certain forms of online speech have effectively sought to prosecute or stifle opposition. Governments may also shut down access to social media sites – or even to the internet – in response to collective action, such as during 2020 anti-government protests in Mali or after Myanmar's 2021 coup.

1.3 Interstate/international influence

International actors involved in conflict abroad often use social media to spread propaganda and disinformation. Such content typically targets any of four audiences: those living in the conflict zone; domestic constituencies; citizens in hostile states; or citizens in neutral or supportive states. During the invasion of Ukraine, a known Russian disinformation operation targeted Facebook and Twitter users in Ukraine's Russian-speaking areas with content portraying their government as corrupt neo-Nazis (Collins and Kent, 2022). Within Russia, civilians have been fed disinformation that called the war a "special military operation" (Sonne and Ilyushina, 2022). False reports of U.S.-run bioweapons labs in Ukraine have been picked up by far-right groups in the United States and Western Europe, as well as by countries more aligned with Russia, including India and China (Chappell and Yousef, 2022). During Azerbaijan's offensive in Nagorno-Karabakh, both Azerbaijan and Armenia used fake accounts to spread propaganda – some of it violent – and harass the other side (Bulos, 2021). These tools are used even when international actors are not a direct party in the conflict. Fake pages and accounts linked to individuals in the French military and Russia's Internet Research Agency (IRA) troll farm competed for influence in the Central African Republic (Graphika and Stanford Internet Observatory, 2020).

Social media may, alternatively, be used to draw international attention to issues related to conflict. During the May 2021 Israel-Palestine escalation, Palestinian activists on social media drew attention to the conflict and encouraged greater solidarity internationally, in part by linking to other protest movements like Black Lives Matter (Yee and El-Naggar, 2021). Ukrainians have used social media during the Russian invasion to push for greater international involvement and to push their narrative of the conflict abroad (Ryan et al., 2022).

1.4 Armed groups

Armed actors often use social media to coordinate, spread propaganda, challenge official narratives, and recruit new members – though their use of platforms often depends on their goals. Syrian rebel groups have used social media platforms to collect donations and recruit fighters (Cohen, 2016;

Berger, 2014). The Arakan Army, an ethnic insurgent group in Myanmar, used Facebook and WhatsApp to issue formal statements, report on fighting with the Tatmadaw, solicit donations, recruit young fighters, and spread propaganda (International Crisis Group, 2021). Hate and terrorist groups use social media to disseminate propaganda: before widespread efforts at deplatforming, ISIS used sites like Twitter and YouTube to glorify its mission, spread violent imagery, and recruit foreign fighters (Siegel and Tucker, 2018; Zeitzoff, 2017). Criminal groups, including in Mexico, use social media to extort kidnapping victims' families, threaten opponents, and recruit new members (International Crisis Group, n.d.). Armed groups often rely on Encrypted Messaging Apps (EMAs) like WhatsApp to communicate amongst themselves.

As in other conflict dimensions, social media can also be important for tracking conflict and reducing support for extremism. In Libya's civil war, rebels relied on social media to publicize human rights violations and provide information about their goals (Jones and Mattiaci, 2017). In Mexico, citizens use anonymous reporting functions to share information about conflict dynamics and violent episodes, with the goal of improving civilian security (Esberg, 2020). Social media has also been used to counter violent extremism, further discussed when reviewing related literature (Briggs and Feve, 2014).

Table 1: Typology of Conflicts

Type	Key Roles for Social Media	Example
Intercommunal Tensions	Hate speech, violence incitement, conflict tracking, counter-messaging, group coordination	Ethiopia; Cameroon; Myanmar; Israel/Palestine; Libya; Sri Lanka; India; Bosnia and Herzegovina; Cyprus
State Repression	Propaganda, suppression, collective action, coordination	Philippines; Myanmar; Venezuela; Mali; Syria; Nicaragua; Belarus
Interstate/International Influence	Propaganda, disinformation, troll farms, international coalition-building	Syria; Central African Republic; Ukraine; Israel-Palestine; Azerbaijan-Armenia;
Armed Groups	Recruitment, extortion, conflict tracking, reducing support	Ethiopia; Syria; Libya; Myanmar; Philippines; Mexico; Nigeria; Colombia

2 Available Data and Limitations within Conflict Zones

Conflict researchers typically use the same set of platform APIs and programs to gather data as those working on other parts of the world do. Indeed, social media can be one of the most accessible forms of data available, particularly in areas with high levels of violence, biased reporting, or media suppression. However, researchers studying conflict face a number of challenges in collecting and interpreting social media data. While these challenges are not always unique to the study of conflict, they are often more consequential for research conducted on unstable, polarized, or developing countries. In particular, we identify two main issues: the interpretation of social media data, which can be biased by issues like internet access and offline conflict events; and limitations to access, which restrict researchers from information related to online and offline behavior. In this section, we outline these issues and their consequences for the use of social media data in conflict zones. After providing an overview of common issues, we then summarize available sources of data and specific limitations across these sources.

2.1 Data Interpretation

2.1.1 Coverage and non-representativeness of user demographics

In interpreting any result from social media, it is vital to understand the demographics of users as well as the actors that may be excluded within the online discourse (Zeitzoff, 2017). Penetration and user characteristics differ greatly across countries, which makes cross-national comparisons difficult. Social media platforms are not randomly generated data sources: for example, Twitter users are likely to be younger and more affluent or more highly educated than the overall population. Who is active on social media platforms is a function of several factors, such as internet access and speed, or access to smartphones (Backer et al., 2016). In addition to demographics, researchers need to be aware of skewed posting rates as well. While some users tweet or post a lot, others only consume content on social media. This has important implications for research: a collection of tweets or posts may not necessarily be a selection of different opinions but rather of repeated opinions from a subsample of the same or similar users. These issues are particularly impactful for conflict zones, because who can access social media, and who posts frequently, is likely a function of the conflict itself.

2.1.2 Unbalanced and lagged reporting

Within conflict zones, certain populations may be more able or more incentivized to produce online content, or their usage may be limited or repressed. As a result, drawing comparisons between groups even within the same country may be challenging. While social media must always be interpreted as a reflection of what users are willing to say publicly, these issues are exacerbated by violence. For

example, repression – or self-censorship out of fear of repression – can significantly change user behavior (Gohdes and Steinert-Threlkeld, 2022). Particularly when using social media as a tool to track violence, researchers should also be aware of delays in reporting violent events in certain areas, due to factors such as displacement, lack of access to the internet, distress, and trauma.

2.1.3 Bias in moderation

Social media companies face a significant challenge when it comes to content moderation, balancing a growing user base and concerns about freedom of expression with the need to limit hateful and dangerous speech. They do so through a combination of automated and human moderation strategies; for the latter, some social media firms hire content moderators (TikTok), while others frequently outsource the work to third-party companies (Facebook, Twitter, YouTube). These moderators, together with AI algorithms, decide on the content that will be taken down versus left on the platform. It has been widely documented that content moderation issues tend to be more severe in non-English speaking environments, and particularly outside the U.S. and western Europe. Social media companies often do not employ a sufficient number of moderators with local language skills – for example, most Arabic-language moderators employed by Facebook speak Moroccan Arabic, and so are less capable of moderating content in Egyptian or Iraqi dialects – and algorithms (including hate speech algorithms) are often less capable at classifying non-English content (e.g. Esberg, 2021). Even moderators with relevant language skills may lack the knowledge of local events to determine which posts violate site rules. Platforms have also been accused of political bias, for example in more heavily censoring Palestinian over Israeli content (HRW, 2021). Moreover, there may be salient differences in the degree to which different conflict parties are willing to flag content – the main way that users can bring potentially harmful content to the attention of moderators. For example, if two groups post the same amount of hate speech that goes undetected by automated flagging methods, but only one side actively reports it, this too may create bias. However, as we discuss later in this section, researchers do not have access to content already removed from platforms.

This creates a challenge for researchers interpreting social media data in conflict areas. For example, take a hypothetical conflict involving two groups, where only one has English as its primary language. If researchers observe an increase in violence incitement from members of the non-English speaking group relative to the English-speaking group, they cannot know whether this is because of an actual relative increase in the posting of such content, or because content moderators are better at catching violating material in English. This can be alleviated by grabbing social media posts in real-time, but often researchers only know what to look for in the aftermath of a significant event. Moreover, the question of bias in content moderation across conflict areas constitutes a significant research topic in and of itself.

2.1.4 Non-human activity and manipulations

Data across platforms can be contaminated by bots and trolls, but also other human efforts to manipulate the algorithms (e.g. on Twitter, using a popular hashtag to promote content that is not related to that hashtag in which case researchers would be extracting data that is less relevant to the research question of their interest). Social media companies are becoming better at spotting fraudulent activities, but the entities creating this content and algorithms are also becoming more sophisticated. Researchers have been working on algorithms to detect fake accounts on Twitter and Facebook (Kagan et al., 2018), but this remains a significant consideration for researchers studying social media. These issues can be particularly problematic in conflict-affected areas, since one side – often the government – typically has better capacity to use bots and trolls.

2.1.5 User-driven changes to available information

The content on social media changes continuously, not only with the addition of new content but also with the changes that users make retrospectively, by deleting posts or accounts. Especially within conflict zones, users may adjust their online behavior and content they produce as a response to the conflict and security-related developments on the ground (Gohdes and Steinert-Threlkeld, 2022). Until recently Reddit maintained the titles of posts deleted by users, as well as any attached comments; however, this appears to be changing in 2022. Other platforms do not provide such metadata. For that reason, researchers need to pay attention to the offline developments and the timing of the data extraction when interpreting the collected data.

2.1.6 Use of Encrypted Messaging Apps (EMAs)

Those living in conflict zones often rely on encrypted services, like WhatsApp and Facebook Messenger, to communicate with one another – especially where there is some physical danger to posting information publicly. This may in effect create a missing data problem. For example, changes in offline context, such as levels of violence, may motivate users to use EMAs instead, which may in turn lead them to using publicly viewable platforms less, or at least differently. As a result, researchers could lack a significant part of the story on the relationship between social media and conflict. While some researchers have used EMAs for experimental interventions, it is of course not possible to gather data on the content of messages. This is particularly concerning for attempts to manage disinformation and violence incitement in content zones, which are often spread on EMAs. For example, during anti-Muslim riots in Sri Lanka, lists of mosques to target spread through WhatsApp (Taub and Fisher 2018).

2.2 Data Access Limitations

2.2.1 Location metadata

Conflict research often focuses on subnational variation in intensity and circumstances. For privacy reasons, however, outside researchers generally cannot access information about where a post was made, or where a user is located. What geolocation data exists is generally user-submitted. Twitter is the most popular platform for researchers relying on geolocation. Prior to 2019, Twitter users could let Twitter geotag their posts, meaning that the coordinates would be extracted from the device that the content was posted from. In 2019, Twitter removed precise geotagging of tweets; the exceptions include opting into sharing location data on a photo taken within the Twitter app or picking a non-precise location tag. The rate of geolocated tweets, however, is low and currently around 1% of all tweets. Researchers have used a spatial label propagation algorithm, based on the locations of accounts' followers, to get user location information (Jurgens, 2013; Jurgens et al., 2015; Mitts, 2019). Instagram cross-posts – which allow for more precise geotagging – have also been exploited (Kruspe et al., 2021).

For obvious reasons, location sharing in conflict-affected areas is particularly dangerous. However, this limits the set of questions that researchers can ask, or requires that they work with only a small subsample of potential posts. In some conflicts, self-reported location may be strategically removed or changed in order to obscure identities, either for safety or for propaganda purposes. Moreover, user-generated posts and reports from the ground circulated on social media can be invaluable to researchers tracking conflict developments. Checking that this information – particularly images and videos – is not fraudulent, however, is a complicated endeavor when Facebook and Twitter generally strip this metadata.

2.2.2 Moderation and taken-down content

Through Facebook's Community Standards, Twitter's Rules, and Instagram's Community Guidelines, social media companies outline what is considered acceptable behavior on their platforms.[2] As described above, while these rules are meant to be applied equally across the globe, there is accumulating evidence that this is not the case. Information previously removed from the platform is not available to researchers, however. This limits the questions that researchers can ask about the spread of hate speech and violence incitement, as well as potentially affecting the tracking of conflict. For example, social media platforms have understandably strict rules around images and videos with extremely graphic content. However, this means that evidence of war crimes may be removed and, as a

[2] See https://research.facebook.com/content-policy-research/; https://help.twitter.com/en/rules-and-policies/twitter-rules; and https://about.instagram.com/blog/announcements/instagram-community-guidelines-faqs.

result, no longer available to researchers, a grave problem for the war in Ukraine and Syria (Clayton, 2022). Obtaining evidence removed from social media has been a challenge for prosecutors, but is also a significant problem for researchers, activists and journalists alike.

2.2.3 Reach, engagement and recommendations

Currently, we have limited understanding of all the ways in which users engage with social media content and an even more limited understanding of what exactly they see on the platforms. The large majority of social media studies rely exclusively on expressions from social media platforms, i.e. information about users' active engagement with the content (posts, likes, shares). We have limited to no information about impressions, i.e. what content users actually viewed and hence either engaged with or chose to skip. Some researchers (Liu et al. 2021) creatively gathered impressions data by running a social media simulation to show that models of influence are incomplete if they only take into account expressions data. Indeed, data on impressions, i.e. exposure, could allow researchers to more rigorously understand the virality of content or spread of misinformation. Doing so is a prerequisite for creating more effective interventions aimed at ameliorating the spread of dangerous rhetoric online.

This would be even more powerful if combined with an enhanced access to data on how users engaged with the content. While researchers can more easily gather engagement data in the form of an overall number of likes, retweets or comments associated with a particular post, there is no accessible data about which posts users actually clicked at or the characteristics of those who engaged with social media content. As a result, researchers have to invest significant amounts of time and resources to make some progress in answering questions that can only be fully and accurately answered with data already available to social media companies.

Moreover, researchers have little ability to study the algorithms that platforms use to promote content. These algorithms are of central importance: recommended videos on YouTube, the content on the top of users' news feeds on Facebook, or the next clip played on TikTok are often how users are exposed to new channels and ideas. Recent evidence suggests that algorithms tend to reward polarizing content, which makes understanding what users see particularly crucial in conflict-affected states (Esberg, 2021). This is especially true for areas where companies do not have sufficient language capacity to automatically flag posts, where changes to algorithms may be the most direct path to reducing the spread of harmful content.

2.2.4 Accessibility

There are significant technical and financial barriers to accessing many social media data sources. Access to the data often requires fairly advanced technical knowledge of R or Python, in addition to the IT infrastructure required to handle larger datasets. There are exceptions to this, most notably Meta's CrowdTangle, which has an easy-to-use interface accessible to journalists, activists, and researchers alike. Access to data often requires permission from the platforms, or certification as a "developer." For most users, Twitter limits access to historical tweets to 7 days. Free access to full historical tweet data is now available (since 2021), yet it should be noted that it is currently only available to researchers on the Academic Research product track. This means that independent researchers or non-profits cannot take full advantage of the Twitter data, though expanding access is identified as one of Twitter's future goals.[3] Even when the data itself can be accessed, a great deal of information comes through images and videos, which require even more significant technical knowledge and infrastructure to process. Most critically, of course, some platforms – notably TikTok – do not allow researchers to access data. As alternative platforms develop, and as existing platforms change, expanding data access will be crucial to understanding the links between social media and conflict.

2.2.5 Additional Source-Specific Limitations

The above outlined a variety of issues related to data interpretation and access that are common to all social media platforms, if to different extents. In Table 2, we provide a general overview of data sources and their specific limitations – particularly when it comes to conflict research.

Table 2: Data Access and Limitations

Source	Description	Platform-Specific Issues
Twitter API	The Sample API is a random sample of all daily Tweets (1% for general users, 10% for those with access to the Twitter Decahose); the Filtered API lets researchers connect to the Twitter "stream" and gather tweets matched to a set of criteria.	Limited access to historical Tweets for regular users (maximum of 7 days); requires knowledge of Python or R; penetration varies a lot by country, but typically biases towards elites.
Twitter Historical	Provides access to any publicly available	Currently available only to academic

[3] See https://blog.twitter.com/developer/en_us/topics/tools/2021/enabling-the-future-of-academic-research-with-the-twitter-api.

PowerTrack and Full-Archive Search	Tweet, starting from March 2006.	researchers; does not provide information about changes to the account over-time; penetration varies a lot by country, but typically biases towards elites.
CrowdTangle	Content discovery tool provided by Meta for easy access to posts from public pages and groups on Facebook, Instagram, and Reddit.	Access is limited to public pages and groups; does not provide access to comments and replies associated with a post.
Meta Ad Library API	Allows for customized keyword searches of ads stored in the Ad Library, allowing researchers to search the data for all active and inactive ads about social issues, elections or politics.	Requires an approved developer account; requires higher level of technical skill; concerns about potential missing data related to more covert political advertising.
Graph API/Marketing API/Pages API	Allows researchers to extract data after registering as a developer.	Data is limited to non-individual public pages; difficult to extract post comments and replies.
Social Science One	Dataset with URLs shared in public posts more than 100 times between 1/1/2017 and 7/31/2019.	Limited timeframe; URL sharing may be less informative for many of the conflict contexts; serious omission detected.
Scrapers for comments/replies	Not a tool made available by social media companies, but can be built by individual researchers.	Difficult to develop and maintain; use may be limited by social media companies.
Meta Data for Good	Launched in 2017, it aggregates data from Meta apps that can be shared in a de-identified way with researchers.	Limited information on humanitarian crises and movement; little access to raw underlying data; limited access to information from sparsely populated or very violent areas; historical data access is limited.
YouTube API	Provides a collection of search results (e.g. video metadata, channel metadata, comments, and closed captions) that a	Requires a fairly high level of technical skill; limited ability for image analysis or text processing in the absence of closed

	researcher sets. By default, projects that enable YouTube data API have a default quota allocation of 10,000 units per day.	captions, which is likely to be the case in content produced in conflict zones.
Reddit API	Researchers can extract data using Reddit API and Python with the pushshift.io API wrapper (which allows the extraction of all Reddit submissions and comments). By default, this returns all historical files for a given query.	Requires a fairly high level of technical skill; Reddit is not a particularly popular resource outside the U.S. and Europe, limiting its usefulness for conflict zones.
TikTok	TikTok does not currently facilitate access to its raw or filtered data.	Lack of data access.
Plug-ins	Researchers can also develop and use different plug-ins for scraping, the most popular including plug-ins for scraping YouTube recommendations, Google search results, or collecting web-browsing data.	Plug-ins may break over time, as the underlying sites change; permissibility of scraping depends on sites' robots.txt.
Social Media Monitoring Products	Companies like Meltwater and Brandwatch allow user-friendly access to a limited selection of social media content, primarily on Twitter.	Expensive; often very limited data samples.

3 Social Media Data in Studies of Conflict

Social media affects and is affected by human behavior and offline socio-political developments. In this section, we focus on existing causal research on topics related to social media and conflict. We classify this research into several categories. First, we summarize literature that treats social media as a key explanatory variable in conflict. Second, we turn to work that considers the impact of offline events on online behavior. Third, we look at a growing literature that seeks to use experimental social media messages to reduce conflict, and particularly intergroup tensions. The last two categories – social media for building conflict data and for the recruitment of experimental subjects – focus on how social media may serve as a valuable tool for conflict research, even when it is not the direct focus of study.

3.1 Social media's effects on conflict

Despite high levels of interest in understanding the effects of social media on conflict, causal research on this topic is difficult: social media may simply reflect, rather than aggravate, offline violence or polarization. Experiments offer one way of overcoming these challenges, through deprivation designs that incentivize users to deactivate from social media platforms for a period of time. These types of designs have been used to explore the effects of social media on civic engagement, psychological wellbeing and polarization. However, they are not suited to areas of active violence, where physical safety could be affected by access to information. The only deprivation RCT conducted within a post-conflict context took place during a week around genocide commemoration in Bosnia and Herzegovina (Asimovic et al., 2021). Participants in this study were recruited through Facebook advertisements, after which they were randomized into treatment (deactivated for a week) and control (remained active on Facebook) groups.[4] Results show that a weeklong deactivation from Facebook worsened users' attitudes toward ethnic outgroups and their knowledge of the news, but improved their subjective well-being. These findings were largely driven by users in more ethnically homogenous offline communities, for whom the online space may provide a unique opportunity to engage – directly or indirectly – with the outgroup. The authors replicated the same design in Cyprus (Asimovic et al., 2022), this time within an ethnically polarized setting with a significant language barrier. Here, however, the authors did not observe the same negative effects of deactivation, raising important questions about the ways in which contextual circumstances (in this case, a language barrier) may shape the effects of social media usage.

While deprivation designs provide individual-level evidence of the effect of social media on conflict, researchers are often interested in its society-wide consequences. Several studies have used the staggered timing of social media access to identify the causal effects of different platforms on political conflict. Fergusson and Molina (2021) exploit variation in the timing of Facebook's release in a given language to study political consequences of Facebook access across a broad set of countries and regions. While their main focus is on the causal impact of social media on protests, they find that Facebook access produces a decrease in violent conflict by deterring violence through increased visibility and serving as a safety valve for voicing discontent. Bursztyn et al. (2019) and Enikolopov et al. (2020) use exogenous variation in Russian social network VK's penetration – based on the particularities of its role out – to show that social media increased hate crimes and protest, respectively. Bursztyn et al. (2019) further show that the rise in hate crimes was the result of platforms facilitating meeting like-minded

[4] Compliance was monitored with an automated Python script that checked users' Facebook URLs, following the procedure from the largest deprivation study to date (Allcott and Gentzkow 2019).

individuals. Warren (2015) provides suggestive, cross-national evidence that higher levels of social media penetration increased collective violence in Africa, particularly in areas lacking other mass media infrastructure.

Researchers have also studied how social media affects conflict developments and intensity. Zeitzoff (2016) used social media data to assess how conflict actors react to changes in public support. Among other collected data, Zeitzoff scraped Twitter feeds of Hamas and Israel's official accounts and several news outlets during the 179 hours of the 2012 conflict. Taking advantage of the fact that supporters were using two competing hashtags (#GazaUnderAttack vs. #IsraelUnderFire) to signal support, he also gathered hashtag data from Twitter firehose. On the combined dataset, Zeitzoff employed Bayesian structural vector autoregression to find that shifts in public support reduce conflict intensity, with the effects being especially strong for Israel. Other research has explored how social media may impact participation in protests against repressive authoritarian regimes: across a variety of contexts, researchers have found that the number of Tweets related to protest-linked keywords increase in the day prior to a major collective action event, suggesting that social media was central for coordination (Acemoglu et al., 2018; Steinert-Threlkeld, 2017; Steinert-Threlkeld et al., 2015; Steinert-Threlkeld and Joo, 2020). Munger et al. (2019), meanwhile, provide evidence that authoritarian regimes also use social media to respond to offline dissent, by showing that pro-regime legislators in Venezuela attempted to distract away from a major national protest movement.

More indirectly, other research has explored how social media affects political speech in conflict. Zhuravskaya et al. (2021) leverage variation in Twitter activity across Israel and Palestine, caused by local Twitter blackouts from lighting strikes or technical failures, to study how social media shapes traditional reporting of conflicts. They find that the content and the tone of U.S. coverage of the Israeli-Palestinian conflict is affected by Twitter: Twitter makes traditional-media reporting of the conflict more emotional and more civilian-oriented, ultimately aiding the side with higher civilian death toll. Mitts et al (2022) use a corpus of 1,072 Islamic State (IS) propaganda videos posted to Twitter from 2015 and 2016 – catching them through the firehose prior to their removal – to explore how propaganda affects online support for extremism. Focusing on the followers of IS accounts, they find that exposure to non-violent propaganda is associated with higher levels of expressed approval for ISIS, while violent imagery decreased support.

3.2 Conflict's effects on social media

While some researchers study social media as the independent variable (i.e. assessing the effects of social media), others focus on how different conflict-related developments influence social media. Establishing causality is again a challenge, given the many factors affecting social media activity. To

overcome issues related to causality, Siegel et al. (2017) collected an original dataset of tweets containing derogatory sectarian slurs posted in 2015 by Twitter users located in the Saudi Kingdom. Exploiting exogenous episodes of foreign violence and domestic mosque attacks, they provide evidence that instances of sectarian violence abroad increase the public expression of anti-Shia content in Saudi Twittersphere. Barcelo and Lazbina (2018) also exploit exogenous timing of terrorist attacks to study the consequences of committing violence on terrorist organizations. They gathered a dataset of 300,842 observations of 13,321 Twitter accounts linked to the Islamic State (IS) by collecting the daily reports made by Anonymous on IS-related Twitter accounts with important information related to the accounts. Merging this dataset with the data on terrorist attacks, they find that IS-terrorist attacks led to a decrease in the number of followers of IS-related Twitter accounts.

Others have exploited social media as a data source that reflects public opinion, which can be particularly difficult to get at through traditional methods in conflict-affected areas due to issues of violence, self-censorship, and state control. Weiss et al. (2021) use 200,000 posts from 71 local Facebook pages and groups, covering 20 municipalities, to show that the announcement of Trump's peace plan in Israel-Palestine was more salient for Palestinian citizens of Israel living in areas where their citizenship would be threatened. Morales (2021) used a corpus of 365,000 Colombian legislators' Tweets, covering 305 politicians, to identify the political effects of FARC attacks. In the short-run, he finds that attacks appear to increase support for more right-wing, "hard-line" Tweets. Barberá et al. (2019) created a dataset of all Twitter and Facebook posts by heads of state between 2012 and 2019, translated into English through natural language processing techniques, then categorized through supervised machine learning methods. Their findings provide evidence that leaders attempt to divert public attention during periods of domestic unrest by emphasizing foreign policy.

Additional research has explored the impacts of state repression on online behavior. Pan and Siegel (2020) use data on the arrests of Saudi dissidents to explore first how the experience of arrests influences their posting on Twitter and, second, how their arrests influence their Twitter followers. They find that, while experiencing arrest tends to quiet government dissent for the person directly affected, it inflames dissent among their followers. Esberg and Siegel (2022) identify the effects of exile on online dissent across Venezuela's opposition: using 5,000,000 Tweets, classified through Word2Vec dictionaries, they find that opponents who fled abroad became more harshly critical of the regime, more supportive of foreign action, and less likely to discuss domestic grievances. In an early working paper, Gohdes and Steinert-Threlkeld (2022) provide evidence that the siege of Aleppo significantly affected Twitter affected users' expression of sentiment, using a difference-in-differences design comparing geolocated tweets from Aleppo to those in other regions.

3.3. Social media for conflict reduction

Understanding strategies for reducing harmful content online is particularly important given the pace with which it can spread on social media, and how powerful it can be in catalyzing violence (Benesch, 2014). These effects are intensified in conflict-affected areas, where online speech has significant potential to translate into offline violence. A number of researchers have thus sought to use social media as a tool to reduce online tensions. Siegel and Badaan (2020) ran an experiment on Twitter during heightened sectarian tensions in Lebanon and a regional proxy war between Iran and Saudi Arabia. They test the effectiveness of different theoretically driven messages (Tweets) in reducing hate speech online, and find elite-endorsed messages priming common religious identity to be most effective.

On Facebook, the most frequent method of delivering experimental treatments takes place via Facebook groups through which researchers embed treatment messages into subjects' feeds. In a recent example of diversifying users' feeds, Scacco et al. (*Working Paper,* 2021) worked with an NGO to randomly expose Jewish Facebook users from Jerusalem to one of 14 posts in their Facebook newsfeeds describing Palestinian life in East Jerusalem. They find that exposure did not shift Jewish attitudes toward Palestinians on aggregate, though there is some evidence that exposure to content highlighting personal Palestinian experiences did improve attitudes. In Cyprus, Asimovic (*Working Paper,* 2022) incentivized users to activate the translation tool on Facebook, in an attempt to reduce language barriers online, and enhance engagement with the content of the outgroup. Preliminary results suggest an improvement in interethnic attitudes among the treated group, with the full study currently being fielded.

While the content of WhatsApp messages is largely inaccessible,[5] some research has begun to deliver experimental treatments on conflict reduction through the platform. In India, Rajeshwari et al. (*Working Paper,* 2022) brought together Hindu and Muslim participants in a series of conversations taking place on WhatsApp (the most popular messaging platform in India). Though the full study is currently being fielded, preliminary evidence from the pilot suggests positive effects of conversations with an outgroup member on outgroup affect and willingness to interact with outgroup members in the future.

3.4 Recruiting participants for experimental research

[5] Some researchers have circumvented this by joining public WhatsApp groups to extract data (e.g. Saha et al., 2021).

When social media is not of substantive interest, it can still serve an important role in collecting experimental data (Guess, 2021). Reaching participants in active conflict zones with traditional recruitment methods often poses an insurmountable challenge. Researchers are rarely able to access conflict areas, either because of the ongoing violence or because of regime hostility to research activities. Even if researchers' physical access was not an obstacle, individuals are unlikely to choose to participate in research using traditional data collection methods due to a myriad of safety concerns.

Recruitment via social media allows investigators to overcome these constraints, although it may mean their samples are not broadly representative. From the researcher's side, the comparatively low cost and speed with which the data can be gathered on social media is particularly advantageous given the constantly changing conflict environment. Another advantage of relying on social media in recruiting subjects is that it allows targeting specific subpopulations of interest (reaching, for example, users who engaged in public discussions about a particular topic; "liked" specific pages or demonstrated other preferences of interest). From participants' perspective, engaging in research through social media may provide them with some agency to decide how much of their identity they want to reveal (compared to, for example, in-person surveys or interviews). Moreover, the flexibility with which they can participate in research – from location to timing – makes such participation less costly than more time-consuming data-collection methods. For example, Williamson and Malik (2020) recruited subjects via Facebook for a survey experiment related to how authoritarian regimes justify repression. Facebook advertisements were explicitly used to help to safely recruit respondents during a time when such research in-country would be dangerous for researchers and participants.

3.5 Developing conflict data

Researchers can also use social media to build datasets *about* conflict, which can be particularly crucial when media coverage and other forms of granular data are unavailable. War, threats against journalists, and lack of free or unbiased media can all make traditional reporting difficult – and make social media a valuable tool for tracking conflict dynamics. For example, recent datasets use visual and text classification methods to identify social media posts about offline collective action (from Won et al., 2017; Zhang and Pan, 2019). Others have used it to better measure violence: Muchlinski et al. (2021) use convolutional neural networks applied to social media text data to develop a new measure of electoral violence, which – because it can more directly link violence to its motives – they claim is 30 percent more accurate at detecting electoral violence than existing datasets. Kotzé et al. (2020) used WhatsApp groups offering eyewitness accounts of crime to build a dataset of violent events in South Africa. Such datasets can also offer a more fine-grained understanding of conflict. Zeitzoff (2011) developed a dataset of hourly dyadic conflict intensity scores from social media (primarily Twitter) to understand conflict dynamics during the Gaza Conflict of 2008-2009. Esberg (2021a) used data

developed from narcoblogs – anonymous citizen journalism pages run through sites like blogger.com – to identify criminal groups in Mexico and where they operate.

Table 3: Social Media and Conflict Research

Research Questions	Examples of Research Designs	Measurement Tools/Capabilities Needed for Future Research
How does social media affect conflict dynamics and other relevant socio-political outcomes?	Deprivation/deactivation designs; natural experiments and difference-in-differences designs (using, e.g., social media blackouts, exogenous variation in social media penetration, staggered rollouts)	Aggregate metrics on content removed from the platform; metrics on reach; (aggregated) location data; information on number of moderators by country/language.
How does conflict affect social media?	Natural experiments and difference-in-differences designs (e.g. leveraging the timing of violent attacks or conflict related events such as peace plan announcements, arrests etc.)	Easier extraction of comments and replies; (aggregated) location data; moderated content data
How can social media be used for conflict reduction?	Interventions encouraging direct contact through social media across geographically segregated groups (e.g. incentivizing intergroup conversations); diversification of users' informational diets (e.g. reducing language barriers through translation settings, diversifying newsfeeds)	Data on engagement and recommendations (to create more effective interventions or target most relevant groups for participation); information on reach.

4 Suggestions

Conflict zones present a number of significant challenges to social media researchers, in terms of both data access and data interpretation. The research presented above worked within these limits to provide insight into the links between social media and conflict – and to use social media to better understand conflict. And some of the challenges are inherent to the use of social media data, including bias in internet access and user demographics. Still, data access issues limit the questions that researchers can ask in a number of important ways. The challenge, however, is how to expand access

while preserving user privacy. This is particularly crucial in violent and polarized contexts, where data mismanagement could have severe consequences. In the following section, we present several recommendations for how social media data may be improved to meet the needs of researchers working in conflict-affected areas.

4.1 Accessing location data

Overall, and particularly within conflict zones, reliability of the location information is low because individuals may hide their location for fear of prosecution and other safety concerns. Protecting users' safety and wellbeing in vulnerable zones has to be the highest priority. Social media companies do have access to location data, via IP address information and video and image metadata. The risks of sharing post-level location with researchers outside these platforms – where users have not explicitly shared this information – is too great, given that such data could easily be misused in conflict-affected areas.

Instead, we propose two possible approaches. First, social media companies could consider sharing location information of where the content was created at different levels of granularity and aggregation – integrating the data to broader administrative levels to better preserve the text content, or aggregating posts according to a set of search parameters to better preserve anonymity. This could allow researchers to study not only events taking place within an area, but also assess how much content about a particular event came from other countries, yielding valuable insights into the influence of foreign actors and audiences on developments within conflict zones. Second, platforms could engage with researchers to identify the accuracy of user-generated location reports, either in their bios or Tweets. For example, platforms may be able to verify using recent IP addresses whether a set of users claiming to live in a particular country actually do, or whether the metadata of particular images or videos were created recently. This could help researchers explore the strategic manipulation of social media in conflict-affected regions, particularly around crucial events like elections, referendums, and significant anniversaries.

4.2. Accessing moderated content

As described above, lack of information about what content has been removed can create significant impediments for researchers. This is particularly true for conflict-affected areas, where bias in moderation due to language or group status may affect what posts can be seen retrospectively. Social media platforms could continue to make moderated content searchable through their APIs or by calling information on specific users, but return only the post time and engagement rather than content. Alternatively, again, social media platforms could share aggregated data on the amount and type of content that was taken down, the areas where content was produced, as well as a set of content

characteristics. Both Facebook and Twitter have increasingly practiced data sharing with research centers like the Stanford Internet Observatory, Graphika, and DFRLab, in order to grant them privileged access to suspected information operation networks prior to removal. Twitter has additionally made the content of a number of these takedowns public. However, platforms should aim to expand this cooperation to allow researchers better access to other forms of harmful content.

Together with this data, researchers should have information on what happens on platforms once they identify harmful content: what percentage of content is removed versus downranked and who is still exposed to downranked content. Within this, researchers could be provided with the statistics around the frequency within which likely non-human activities, third-party application activities and manipulation attempts were detected. Finally, given the accumulated evidence on moderation bias, the information about how many moderators are employed for each region and how well the algorithms involved in the moderation process are performing across languages should be more easily accessible. Without clear information about the extent of asymmetries when it comes to the investment in content moderation, researchers may be making erroneous conclusions in comparing data across groups and regions.

Conflict zones, however, face a unique challenge in the moderation of content. For good reason, most social media platforms heavily moderate violent imagery, often removing images and videos deemed too graphic. Yet this publicly-posted content is often the best documentation available of human rights violations and war crimes. Human Rights Watch has criticized social media platforms for failing to ensure that the content removed is preserved, archived, and made available for prosecutors. Some have called for a more centralized system of uploads from conflict zones – so called "digital lockers" – but there has been no success in making this a reality so far (HRW, 2021a; Human Rights Center, 2021).

Finally, this is a space of continuous changes: most recently (April 2022), the EU reached a deal on the legislation titled Digital Services Act, aimed at addressing illegal and harmful content, transparent advertising and misinformation on social media platforms. Among other regulations, the DSA would effectively prevent targeting of users with algorithms that use the data based on gender, race and religion; and governments would be able to request social media platforms to take down material deemed illegal such as hate speech, incitement to terrorism and child sexual abuse. Importantly, tech companies would also be asked to carry out annual independent audits and follow transparency requirements for content ranking algorithms, while the Commission will have the power to impose fines of up to 6% of platforms global revenue for non-compliance.

4.3 Improving accessibility

While we welcome the new developments in the realm of social media data collection, we hope that companies continue to build tools that would also serve the needs of those who may not have the programming training, technical infrastructure, or available funds required for studying social media. These issues are often particularly acute in areas in conflict-affected areas, meaning that those most affected by conflict may face the most challenges in gaining access to data. In addition to more user-friendly tools to download data, which CrowdTangle is a model of, this could include integrating basic visualization tools, particularly helpful where technical infrastructure may not support analyzing large amounts of raw data.

4.4 Data on impressions, engagement and recommendations

Providing researchers access to data on impressions can significantly expand the questions available to researchers working on conflict-affected areas. Conflict zones often see rumors develop that could lead to violence, but are not deemed violence incitement – for example, a recent "unverified rumor" posted to Facebook in Ethiopia claiming that Tigray civilians were participating in atrocities against the Amhara people.[6] Understanding whether the spread of these posts is due to user behavior – seeing both factual information and rumor, but only engaging with the rumor – or whether these rumors tend to be more widely viewed would aid in efforts to reduce the visibility of such content. Researchers also need enhanced access to richer data engagement data, in particular data about which posts users actually clicked at and the characteristics of those who engaged with social media content (e.g. demographics breakdown of those who reacted to social media posts).

Similarly, while algorithms may be highly tailored to the individual, providing information about the most commonly recommendations on a YouTube video or Facebook group can help understand how content spreads online. This applies but is not limited to Twitter, Facebook, YouTube, Instagram, and Reddit. WhatsApp could provide aggregated information about number and size of public groups, and the frequency of messages that are circulated during aggregated periods of times.

4.5 Expanding data features

As described in Table 2, platforms often have specific limitations on data access. We want to highlight two issues that particularly affect researchers of conflict on the largest social media platforms. First, Facebook does not currently allow researchers to extract comments and replies through CrowdTangle, and it is difficult to extract this information even through the API. This is, however, where most

[6] See https://transparency.fb.com/oversight/oversight-board-cases/raya-kobo-ethiopia.

discussions happen and is important for understanding conflict dynamics on social media, particularly since access is limited to public groups and pages. Adding a feature to CrowdTangle that would extract de-anonymized comments and replies would allow for a richer analysis of the social media ecosystem.

Second, most metadata associated with a Tweet is extracted at the time of data collection, rather than the Tweet creation time. This means that information relating to follower networks as well as other profile information, such as the account name or bio, will be from the moment of data collection. This limits the types of questions researchers can ask: for example, studying how changing elite rhetoric or conflict intensity affects who users choose to (un)follow or how they choose to identify in their bios. Allowing researchers to access information about users' follower networks or bios at the Tweet creation time would allow for a richer set of questions to be analyzed.

Conclusion

In this report, we identified the types of conflicts where social media plays a role, described social media data that are available to researchers, and provided examples of how this data has been used in academic research in conflict-affected contexts. We also identified main challenges related to extraction and interpretation of social media data, which provided basis for our suggestions around expanding data access.

Recent years have clearly shown how large of an influence social media can have on information ecosystems and political developments within conflict-affected areas. As the number of users within these contexts continues to grow, the need to equip independent researchers with enhanced data access and more transparency on platform's decision-making processes has never been more pressing. We welcome some of the more recent efforts to strengthen partnerships between social media and academia. The current system, however, continues to fall short of what is needed for researchers to effectively provide insights that would inform strategies on how social media can be a tool of service in conflict-affected areas, rather than a tool of further destruction. For conflict-affected zones – where time and resources are particularly scarce – the current inefficiencies and limitations around data collections can bring grave consequences. Overcoming these limitations in a way that protects the privacy and safety of those in adverse circumstances will require creative thinking, close collaboration between researchers, social media companies and relevant stakeholders, and a dedication to serving the communities that need them.

References

Acemoglu, Daron, Tarek A. Hassan, and Ahmed Tahoun. 2017. "The Power of the Street: Evidence from Egypt's Arab Spring." *The Review of Financial Studies* 31(1): 1–42.

Alami, Mona. 2018. "Russia's Disinformation Campaign Has Changed How We See Syria." *Atlantic Council*. https://www.atlanticcouncil.org/blogs/syriasource/russia-s-disinformation-campaign-has-changed-how-we-see-syria/ (May 9, 2022).

Alba, Davey. 2018. "How Duterte Used Facebook to Fuel the Philippine Drug War." *BuzzFeed News*. https://www.buzzfeednews.com/article/daveyalba/facebook-philippines-dutertes-drug-war.

Asimovic, Nejla, Jonathan Nagler, Richard Bonneau, and Joshua A. Tucker. 2021. "Testing the Effects of Facebook Usage in an Ethnically Polarized Setting." *Proceedings of the National Academy of Sciences* 118(25).

Barberá, Pablo, Anita R. Gohdes, Evgeniia Iakhnis, and Thomas Zeitzoff. 2019. "Distract and Divert: How World Leaders Use Social Media during Contentious Politics." Working Paper.

Barceló, J., & Labzina, E. (2020). Do Islamic state's deadly attacks disengage, deter, or mobilize supporters?. *British Journal of Political Science*, 50(4), 1539-1559.

Benesch, Susan. 2014. *Defining and Diminishing Hate Speech*. Minority Rights Group International. https://www.minorityrights.org/wp-content/uploads/old-site-downloads/mrg-state-of-the-worlds-minorities-2014-chapter02.pdf.

Briggs, Rachel, and Sebastien Feve. 2014. *Countering the Appeal of Extremism Online*. Institute for Strategic Dialogue https://www.dhs.gov/sites/default/files/publications/Countering%20the%20Appeal%20of%20Extremism%20Online_1.pdf.

Bulos, Nabih. 2020. "Around Nagorno-Karabakh, an All-out Media War Unfolds." *Los Angeles Times*. https://www.latimes.com/world-nation/story/2020-11-10/nagorno-karabakh-propaganda-war-armenia-azerbaijan (November 20, 2020).

Bursztyn, Leonardo, Georgy Egorov, Ruben Enikolopov, and Maria Petrova. 2019. "Social Media and Xenophobia: Evidence from Russia." *NBER*. https://www.nber.org/papers/w26567.

Chappell, Bill, and Odette Yousef. 2022. "How the False Russian Biolab Story Came to Circulate among the U.S. Far Right." *NPR*. https://www.npr.org/2022/03/25/1087910880/biological-weapons-far-right-russia-ukraine. (May 10, 2022).

Cheney, Catherine. 2022. "Ukraine Shows Promise and Peril of Satellite Internet in War Zones." *Inside Development*. https://www.devex.com/news/ukraine-shows-promise-and-peril-of-satellite-internet-in-war-zones-102794..

Chulov, Martin. 2020. "How Syria's Disinformation Wars Destroyed the Co-Founder of the White Helmets." *The Guardian*. https://www.theguardian.com/news/2020/oct/27/syria-disinformation-war-white-helmets-mayday-rescue-james-le-mesurier.

Clayton, James. 2022. "Are Tech Companies Removing Evidence of War Crimes?" *BBC News*. https://www.bbc.com/news/technology-60911099 (May 10, 2022).

Collins, Ben, and Jo Ling Kent. 2022. "Facebook, Twitter Remove Disinformation Accounts Targeting Ukrainians." *NBC News*. https://www.nbcnews.com/tech/internet/facebook-twitter-remove-disinformation-accounts-targeting-ukrainians-rcna17880.

"Crime Online: How Mexico's Criminal Groups Use Social Media." 2022. *Working report, International Crisis Group*.

Digital Lockers: Archiving Social Media Evidence of Atrocity Crimes. 2021. Human Rights Center, UC Berkeley School of Law. https://humanrights.berkeley.edu/sites/default/files/digital_lockers_report5.pdf (May 10, 2022).

"Easing Cameroon's Ethno-Political Tensions, on and Offline." 2020. *International Crisis Group*. https://www.crisisgroup.org/africa/central-africa/cameroon/295-easing-cameroons-ethno-political-tensions-and-offline.

Enikolopov, Ruben, Alexey Makarin, and Maria Petrova. 2020. "Social Media and Protest Participation: Evidence from Russia." *Econometrica* 88(4): 1479–1514.

Esberg, Jane. 2021a. "Licit Commodities and Criminal Violence." *Working Paper*.

———. 2021b. "What the Facebook Whistleblower Reveals about Social Media and Conflict." *International Crisis Group*. https://www.crisisgroup.org/united-states/united-states-internal/what-facebook-whistleblower-reveals-about-social-media-and-conflict.

Esberg, Jane, and Alexandra Siegel. 2022. "How Exile Shapes Online Opposition: Evidence from Venezuela." *IPL Working Paper Series* 21(1).

Fergusson, Leopoldo, and Carlos Molina. 2019. "Facebook Causes Protests." *Working Paper*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3553514 (May 10, 2022).

Frenkel, Sheera. 2021. "Lies on Social Media Inflame Israeli-Palestinian Conflict." *The New York Times*. https://www.nytimes.com/2021/05/14/technology/israel-palestine-misinformation-lies-social-media.html.

Gohdes, Anita, and Zachary C. Steinert-Threlkeld. 2022. "Civilian Signalling on Social Media during Civil War." *Working Paper*.

Guo, Eileen, and Patrick Howell O'Neill. 2021. "Millions of People Rely on Facebook to Get Online. The Outage Left Them Stranded." *MIT Technology Review*. https://www.technologyreview.com/2021/10/05/1036479/facebook-global-outage/#:~:text=For%20much%20of%20the%20world.

Guriev, Sergei, Nikita Melnikov, and Ekaterina Zhuravskaya. 2020. "3G Internet and Confidence in Government." *The Quarterly Journal of Economics* 136(4): 2533–2613.

"Hate Speech in Ethiopia: Abiy Ahmed Resurrects Old Demons." 2020. *DW*. https://www.dw.com/en/hate-speech-in-ethiopia-abiy-ahmed-resurrects-old-demons/a-55800705.

"Israel/Palestine: Facebook Censors Discussion of Rights Issues." 2021. *Human Rights Watch*. https://www.hrw.org/news/2021/10/08/israel/palestine-facebook-censors-discussion-rights-issues# (May 10, 2022).

Jones, Benjamin T., and Eleonora Mattiacci. 2017. "A Manifesto, in 140 Characters or Fewer: Social Media as a Tool of Rebel Diplomacy." *British Journal of Political Science* 49(2): 739–61.

Jurgens, David. 2013. "That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships." *Proceedings of the International AAAI Conference on Web and Social Media* 7(1): 273–82.

———. 2015. "Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice." *Proceedings of the International AAAI Conference on Web and Social Media* 9(1): 188–97.

Kotzé, Eduan, Burgert A. Senekal, and Walter Daelemans. 2020. "Automatic Classification of Social Media Reports on Violent Incidents in South Africa Using Machine Learning." *South African Journal of Science* 116(3/4).

Kruspe, Anna et al. 2021. "Changes in Twitter Geolocations: Insights and Suggestions for Future Usage." *Working paper*. https://arxiv.org/abs/2108.12251.

Liu, Rui et al. 2021. "Using Impression Data to Improve Models of Online Social Influence." *Scientific Reports* 11(1).

Majumdar, Rajeshwari, Richard Bonneau, Jonathan Nagler, Joshua A. Tucker. "Reducing Intergroup Prejudice through Conversations on WhatsApp." Working Paper.

Mitts, Tamar. 2018. "From Isolation to Radicalization: Anti-Muslim Hostility and Support for ISIS in the West." *American Political Science Review* 113(1): 173–94.

Mitts, Tamar, Gregoire Phillips, and Barbara Walter. 2021. "Studying the Impact of ISIS Propaganda Campaigns." *The Journal of Politics* 84(2).

Morales, Juan S. 2021. "Legislating during War: Conflict and Politics in Colombia." *Journal of Public Economics* 193.

More-Troll Kombat. 2020. Graphika and Stanford Internet Observatory. https://public-assets.graphika.com/reports/graphika_stanford_report_more_troll_kombat.pdf (May 10, 2022).

Mozur, Paul. 2018. "A Genocide Incited on Facebook, with Posts from Myanmar's Military." *The New York Times*. https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html.

Muchlinski, David et al. 2020. "We Need to Go Deeper: Measuring Electoral Violence Using Convolutional Neural Networks and Social Media." *Political Science Research and Methods* 9(1): 1–18.

Munger, Kevin, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. 2018. "Elites Tweet to Get Feet off the Streets: Measuring Regime Social Media Strategies during Protest." *Political Science Research and Methods* 7(04): 815–34.

"Myanmar's Military Struggles to Control the Virtual Battlefield." 2021. *Crisis Group*. https://www.crisisgroup.org/asia/south-east-asia/myanmar/314-myanmars-military-struggles-control-virtual-battlefield (May 10, 2022).

Pan, Jennifer, and Alexandra A. Siegel. 2019. "How Saudi Crackdowns Fail to Silence Online Dissent." *American Political Science Review* 114(1): 109–25.

"Philippines: End Deadly 'Red-Tagging' of Activists." 2022. *Human Rights Watch*. https://www.hrw.org/news/2022/01/17/philippines-end-deadly-red-tagging-activists.

Ryan, Missy, Ellen Nakashima, Michael Birnbaum, and David L. Stern. 2022. "Outmatched in Military Might, Ukraine Has Excelled in the Information War." *Washington Post*. https://www.washingtonpost.com/national-security/2022/03/16/ukraine-zelensky-information-war/.

Saha, Punyajoy, Binny Mathew, Kiran Garimella, and Animesh Mukherjee. 2021. "'Short Is the Road That Leads from Fear to Hate': Fear Speech in Indian WhatsApp Groups." *Proceedings of the Web Conference 2021*.

Scacco, Alexandra, Alexandra A. Siegel, and Chagai M. Weiss. 2021. "Desegregating Digital Spaces: A Facebook Field Experiment in Jerusalem." Working Paper. https://www.wzb.eu/system/files/docs/ped/ipi/Desegregating_Digital_Spaces_SSW_0.pdf.

Siegel, Alexandra A., and Vivienne Badaan. 2020. "#No2Sectarianism: Experimental Approaches to Reducing Sectarian Hate Speech Online." *American Political Science Review* 114(3): 837–55.

Siegel, Alexandra A., and Joshua A. Tucker. 2017. "The Islamic State's Information Warfare." *Journal of Language and Politics* 17(2): 258–80.

Siegel, Alexandra, Joshua Tucker, Jonathan Nagler, and Richard Bonneau. 2017. "Socially Mediated Sectarianism: Violence, Elites, and Anti-Shia Hostility in Saudi Arabia." Working paper. https://alexandra-siegel.com/wp-content/uploads/2019/05/Siegel_Sectarianism_January2017.pdf (May 10, 2022).

Sobolev, Anton, M. Keith Chen, Jungseok Joo, and Zachary C. Steinert-Threlkeld. 2020. "News and Geolocated Social Media Accurately Measure Protest Size Variation." *American Political Science Review* 114(4): 1343–51. https://www.anderson.ucla.edu/faculty_pages/keith.chen/papers/WP_MeasuringProtestSize.pdf (September 13, 2021).

Sonne, Paul, and Mary Ilyushina. 2022. "Inside Russia's Propaganda Bubble: Where a War Isn't a War." *Washington Post*. https://www.washingtonpost.com/world/2022/03/17/russia-information-firewall-ukraine-war/.

Steinert-Threlkeld, Zachary C, Delia Mocanu, Alessandro Vespignani, and James Fowler. 2015. "Online Social Networks and Offline Protest." *EPJ Data Science* 4(1). https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-015-0056-y (May 10, 2022).

Steinert-Threlkeld, Zachary C. 2017. "Spontaneous Collective Action: Peripheral Mobilization during the Arab Spring." *American Political Science Review* 111(2): 379–403.

Stevenson, Alexandra. 2018. "Facebook Admits It Was Used to Incite Violence in Myanmar." *The New York Times*. https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html.

Suárez Pérez, Daniel, Esteban Ponce, and Leon Rosas. 2021. *DIGITAL AUTOCRACY Maduro's Control of the Venezuelan Information Environment*. DFRLab. https://www.atlanticcouncil.org/wp-content/uploads/2021/04/DigitalAutocracyVEN-FINAL.pdf (May 10, 2022).

Taub, Amanda, and Max Fisher. 2018. "Where Countries Are Tinderboxes and Facebook Is a Match." *The New York Times*. https://www.nytimes.com/2018/04/21/world/asia/facebook-sri-lanka-riots.html.

"Video Unavailable": Social Media Platforms Remove Evidence of War Crimes. 2021. Human Rights Watch. https://www.hrw.org/sites/default/files/media_2020/09/crisis_conflict0920_web_0.pdf (May 10, 2022).

Warren, T Camber. 2015. "Explosive Connections? Mass Media, Social Media, and the Geography of Collective Violence in African States." *Journal of Peace Research* 52(3): 297–311.

Weiss, Chagai, Alexandra Siegel, and David Romney. "How Threats of Exclusion Mobilize Palestinian Political Participation." *American Journal of Political Science* Forthcoming.

Williamson, Scott, and Mashail Malik. 2020. "Contesting Narratives of Repression: Experimental Evidence from Sisi's Egypt." *Journal of Peace Research* 58(5): 1018–33.

Won, Donghyeon, Zachary C. Steinert-Threlkeld, and Jungseock Joo. 2017. "Protest Activity Detection and Perceived Violence Estimation from Social Media Images." *Proceedings of the 25th ACM international conference on Multimedia*.

Yee, Vivian, and Mona El-Naggar. 2021. "'Social Media Is the Mass Protest': Solidarity with Palestinians Grows Online." *The New York Times*. https://www.nytimes.com/2021/05/18/world/middleeast/palestinians-social-media.html.

Zeitzoff, Thomas. 2011. "Using Social Media to Measure Conflict Dynamics." *Journal of Conflict Resolution* 55(6): 938–69.

———. 2016. "Does Social Media Influence Conflict? Evidence from the 2012 Gaza Conflict." *Journal of Conflict Resolution* 62(1): 29–63.

———. 2017. "How Social Media Is Changing Conflict." *Journal of Conflict Resolution* 61(9): 1970–91.
Zhang, Han, and Jennifer Pan. 2019. "CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media." *Sociological Methodology* 49(1): 1–57.

Appendix

A.1 Codebook

Codebook for review of collections and analytics useful for understanding role of social media in conflict zones

Article Inclusion Criteria

Articles for inclusion were identified in two stages. First, we searched for published articles and working papers listed on Google Scholar from a search of "social media" plus "protest," "violence," "conflict," "war," "polarization," or "post conflict." We then reviewed the abstracts for relevancy, to filter for only research meeting the following criteria:

- Draws on data developed from social media platforms;
- Uses causal quantitative methods;
- Focuses on conflict-affected or post-conflict countries, e.g. those countries that currently or in the past experienced significant deadly violence

We do not limit ourselves by field, since some relevant work comes from fields outside of the social sciences.

Second, we use information on these articles to identify related work. Relevant content can be difficult to find due to a large number of unrelated articles (e.g., social media and "workplace conflict"). For publications identified as relevant, we review both the works cited and work that cites the article, via Google Scholar. If we find multiple relevant articles for a single author (e.g. Thomas Zeitzoff or Alexandra Siegel), we additionally review their CVs for additional content, including working papers. We have summarized some of these papers within the main text, and selected a few that are described in more detail within the attached "Research-Sources" document. These papers were selected to reflect

the diversity of research questions, methods and conflict settings that researchers have been focusing on over the recent years.

1. Data Fields

1. Article
    1.1. Bibliographic data (note – will be slightly different for conference proceedings)
        1.1.1. Authors
        1.1.2. Title
        1.1.3. Journal
        1.1.4. Year
        1.1.5. Volume
        1.1.6. Number
        1.1.7. Pages
        1.1.8. Full .bib citation
    1.2. Abstract

2. Level of analysis (What kind of dataset is wrangled?)
    2.1. User-level (does the dataset have users as the unit of analysis?)
        2.1.1. Aggregate content by period
        2.1.2. User-level platform behavior
        2.1.3. Survey data
        2.1.4. Non-human (bots) account-level data
        2.1.5. Network behavior at user level (following, being followed, linking)
        2.1.6. Other (describe)
    2.2. Post Level (For social media content, does the analysis happen on posts/tweets/videos, etc.)
        2.2.1. User Posts
        2.2.2. Direct Interaction with(within) the posts - Reactions/ Comments
        2.2.3. Indirect interaction - Shares/ Retweets
        2.2.4. Other (describe)
    2.3. [For Post-Level] Primary Content Type
        2.3.1. Text
        2.3.2. Images
        2.3.3. Video
        2.3.4. Hyperlinks
        2.3.5. Audio
    2.4. Network-level (Any social network constructed or created for the analysis?)

2.5. Custom Aggregations
    2.5.1. Platform (Period)
    2.5.2. Platform (Location)
    2.5.3. Other aggregation (describe)
2.6. Time-aggregation (Was the dataset aggregated on any of these time parameters?)
    2.6.1. Hourly
    2.6.2. Daily
    2.6.3. Weekly
    2.6.4. Monthly
    2.6.5. Yearly
    2.6.6. Quarterly/ Bunched months
    2.6.7. Other (describe)
2.7. Cross-platform
    2.7.1. Follow users across platforms
    2.7.2. Follow content across platforms
        2.7.2.1. Text
        2.7.2.2. Images
        2.7.2.3. Video
        2.7.2.4. Hyperlinks

3. Research design
    3.1. Key question
    3.2. Countries studied
    3.3. Country selection (why countries were chosen)
    3.4. Primary outcome
    3.5. Primary treatment (if applicable)
    3.6. Type of study/Method of analysis
        3.6.1. Experimental
        3.6.2. Causal
            3.6.2.1. Difference-in-Differences
            3.6.2.2. Natural Experiment
        3.6.3. Machine Learning
            3.6.3.1. Supervised
            3.6.3.2. Unsupervised
            3.6.3.3. Semi-supervised
        3.6.4. Network Analysis
        3.6.5. Other (Describe)

4. Data Sources. The following fields are completed for each article*source pair. If an article uses both Facebook and Twitter data, for example, it will have two rows in the dataset.
    4.1. Platform (e-mail, Twitter, Facebook)

4.2. Date range
4.3. Observation level (post, user/day, etc.)
4.4. # of observations
4.5. Sampling/selection strategy
    4.5.1. On demographics
    4.5.2. Sampling content from a fixed period
    4.5.3. Selecting users who interact with a particular content
    4.5.4. Influencers
    4.5.5. Keyword based selections
    4.5.6. Others (describe)
4.6. Collection method
    4.6.1. Scraping
    4.6.2. API (e.g. CrowdTangle or Twitter Research API)
    4.6.3. Company data release
    4.6.4. Independent collections
    4.6.5. Institutional partnerships
    4.6.6. Social Listening Tool + Bots (created by authors)
    4.6.7. Browser extensions
    4.6.8. Crowdsourced/ Survey data
    4.6.9. Manual annotations or coding
    4.6.10. Others (describe)
4.7. Data Type
    4.7.1. Realtime
    4.7.2. Retrospective
    4.7.3. Both

5. Features extracted (if any) post analysis
    5.1. NLP
        5.1.1. Standardized dictionary application (e.g. LIWC)
        5.1.2. Supervised ML of custom labels
        5.1.3. Unsupervised ML and applying labels (LDA topic models)
    5.2. Network Features
    5.3. Others
        5.3.1. Results from causal-inference techniques