RTI Press

Occasional Paper

ISSN 2378-7996

May 2023

Identifying and Classifying COVID-19 Stigma on Social Media

Nancy Warren, Pia Mingkwan, Caroline Kery, Meagan Meekins, Thomas Bukowski, and Laura Nyblade

RTI INTERNATIONAL

RTI Press publication OP-0087-2305

RTI International is an independent, nonprofit research organization dedicated to improving the human condition. The RTI Press mission is to disseminate information about RTI research, analytic tools, and technical expertise to a national and international audience. RTI Press publications are peer-reviewed by at least two independent substantive experts and one or more Press editors.

Suggested Citation

Warren, N., Mingkwan, P., Kery, C., Meekins, M., Bukowski, T., Nyblade, L. (2023). *Identifying and Classifying COVID-19 Stigma on Social Media*. RTI Press Publication No. OP-0087-2305. Research Triangle Park, NC: RTI Press. https://doi.org/10.3768/rtipress.2023.op.0087.2305

This publication is part of the RTI Press Research Report series. Occasional Papers are scholarly essays on policy, methods, or other topics relevant to RTI areas of research or technical focus.

RTI International
3040 East Cornwallis Road
PO Box 12194
Research Triangle Park, NC
27709-2194 USA

Tel: +1.919.541.6000
E-mail: rtipress@rti.org
Website: www.rti.org

Contents

About the Authors

Nancy Warren,* MPH, is a program manager in the Global Health Division at RTI International. https://orcid.org/0000-0003-3226-074X

Pia Mingkwan,* MSc, was at RTI International in the Global Health Division at the time of the research. https://orcid.org/0000-0001-5746-6096

Caroline Kery, MS, is a research data scientist at the Center for Data Science at RTI International. https://orcid.org/0000-0002-1794-3369

Meagan Meekins, MPH, is a learning and knowledge management facilitator for the International Development Group at RTI International. https://orcid.org/0000-0003-0207-9667

Thomas Bukowski, MPA, MA, is a digital and social media health research analyst for the Center for Health Analytics, Media, and Policy's Health Media Impact and Digital Analytics program at RTI International. https://orcid.org/0000-0001-5720-8564

Laura Nyblade, MA, PhD, is a Fellow in Health Policy and expert in stigma and discrimination in the Global Public Health Impact Center at RTI International. https://orcid.org/0000-0001-8067-4192

*These authors contributed equally as co-first authors.

RTI Press Associate Editor

Pia MacDonald

Acknowledgments

The research team would like to acknowledge the Scientific Stature and Research Committee and leadership of the International Development Group at RTI International for its prioritization of funds to conduct this important work. NW and PM contributed equally to this work as co-first authors.

Abstract

Since the introduction of COVID-19 in early 2020, COVID-19 stigma has persisted on social media. Stigma, a social process where individuals or groups are labeled, stereotyped, and separated, can result in misinformation, discrimination, and violence. The body of research on COVID-19 stigma is growing, but addressing stigma on social media remains challenging because of the enormous volume and diversity of rapidly changing content. This three-part methodology offers a standardized approach for generating (1) a relevant and manageable social media sample for stigma identification and research, (2) a categorization process to organize the sample, and (3) a systematic coding method for classifying stigma within the sample. An application of the methodology generated a curated sample of 138,998 posts from Twitter and Reddit, organized according to key stigma domain, key terms, frequency of terms, and hashtag occurrence. A subset of 711 posts were selected for the content analysis and analyzed based on the key stigma domains, distinguishing between intentional and unintentional stigma. This methodology has the potential to facilitate comprehensive social media stigma research through simplified sample generation and stigma identification processes and offers the possibility of adaptation to address other types of social media stigma, beyond COVID-19.

Introduction

> "…we had to find a name that did not refer to a geographical location, an animal, an individual or group of people, and which is also pronounceable and related to the disease. Having a name matters to prevent the use of other names that can be inaccurate or stigmatizing. It also gives us a standard format to use for any future coronavirus outbreaks."
>
> WHO Director-General Tedros Adhanom Ghebreyesus, February 11, 2020

In February 2020, World Health Organization (WHO) Director-General Dr. Tedros Adhanom Ghebreyesus tweeted the following message: "Our greatest enemy right now is not the coronavirus itself; it is fear, rumor and **stigma**. Our greatest assets are facts, reason and solidarity" (WHO, 2020a). As new variants of COVID-19 have continued to emerge and as countries have struggled to adapt effective plans of action, the need to address COVID-19–related stigma remains urgent.

COVID-19 stigma occurs when people are "labelled, stereotyped, separated, and/or experience loss of status and discrimination because of a potential negative affiliation with the disease" (World Health Organization (WHO) et al., 2020). Since February 2020, reports of COVID-19–related stigma and discrimination have been persistent, particularly among specific groups such as COVID-19 patients and survivors; immigrants; people of Asian descent; and populations that are already stigmatized for health or lifestyle conditions (e.g., smokers, people living with HIV, members of the LGBTQ+ community) (Bloomberg, 2020; Jan, 2020; Krishnatray, 2020; WHO, 2020b; WHO et al., 2020). Previous research, globally and by RTI International's experts, has established the relationship between misinformation and stigma and the harm it can have on health outcomes and access to care (Asadi-Aliabadi et al., 2020; Hatzenbuehler et al., 2013; Kim et al., 2018; Nyblade et al., 2019). Experiences from other epidemics, such as HIV, SARS, and Ebola, have emphasized the impact of infectious disease stigma on testing, treatment, recovery, and other health behaviors (Churcher, 2013; Ekstrand et al., 2018; Gesesew et al., 2017; Gourlay et al., 2013; Hamilton et al., 2019; Kim et al., 2018; Logie, 2020; Nyblade et al., 2019; Rueda et al., 2016; Turan & Nyblade, 2013). Current COVID-19 stigma mitigation strategies are informed by over two decades of infectious disease stigma research (DuPont-Reyes et al., 2020; Logie, 2020; Yuan et al., 2021) but lack a comprehensive approach to identify COVID-19 stigma in all of its manifestations, including in social media content.

The rapid dissemination of information about COVID-19 via social media has resulted in increased availability of information and has offered a platform for users to express themselves and ask questions. As a result, the large volume of COVID-19–related messaging on social media poses a challenge to identifying, managing, and addressing stigmatizing content. The largely unregulated sharing of opinions and information about COVID-19 on social media has also increased the risk of misleading and misinforming users, perpetuating the fear, ignorance, or xenophobia that contributes to COVID-19 stigma (Azim et al., 2020; Wang et al., 2019).

Successful stigma reduction interventions, like those created to address HIV stigma and discrimination, have focused on addressing the key actionable drivers of stigma, which include awareness, fear/discomfort, attitudes, and institutional environments (Earnshaw & Chaudoir, 2009; Nyblade et al., 2021; Nyblade et al., 2019; Stangl et al., 2019). These successful methodologies emphasize that a strong understanding of and ability to identify and classify stigma is essential to combatting it. Current research has used these existing global best practices to begin to name and address COVID-19 stigma (Asadi-Aliabadi et al., 2020; Huda et al., 2020; Krishnatray, 2020; Logie, 2020; WHO et al., 2020; Yuan et al., 2021) but little evidence exists on how to apply these stigma measurement tools to social media research, particularly the unique data types, volume, and creation frequency of social media content. Efforts to combat COVID-19 stigma on social media must be informed by a strong understanding of how COVID-19 stigma manifests across a sample size as large as millions of social media posts. A strong, standardized approach for identifying and classifying COVID-19 stigma on social media will offer future researchers and policy makers the ability to identify

it on a massive scale; understand how it manifests; and ultimately develop policies, systems, and interventions to combat and prevent it.

The goal of this research was to develop a comprehensive methodology for identification of COVID-19 stigma on social media by adapting current best practices for stigma identification and measurement for the scope and variety of social media data. This exploratory research proposes a methodology to identify and classify COVID-19 stigmatizing content on social media using a key terms matching system correlated to evidence-based definitions of stigma. The methodology offers a standardized approach for (1) generating a relevant and manageable sample of content from social media for stigma identification and research, (2) a categorization process to organize the sample, and (3) a systematic coding method for classifying stigma within the sample.

The methodology was developed using data from Twitter and Reddit, two popular and well-established platforms. Twitter, a microblogging and social networking service, facilitates quick, frequent communication through short posts titled "tweets," which may contain photos, videos, links, and text. All posts are tied to a user, generally available to their followers, and can be "retweeted" by other users.[1] Posts are also searchable, and posts designated as public are widely searchable through third-party search engines. Reddit is a slightly different type of social networking service, one that is community-based rather than user-centric, and comprises social news aggregation, discussion, and content rating. The site comprises sub-communities, or "subreddits," where users share and discuss different topics. These two platforms were chosen specifically for this research because they have open-access application programming interfaces (APIs[2]) and because both have millions of users and data points (in terms of daily active users, Twitter has 238 million and Reddit has over 50 million) (Reddit Inc., 2022; Twitter, 2022). Unfortunately, the specific demographics of the platforms' users are not readily available, but the breadth and length of the sampling strategy used in this research was designed to promote a representative sample of relevant content. Other well-known platforms such as Facebook or Instagram have significant active user numbers but also have much stricter data access regulations and closed APIs, making it nearly impossible to analyze the same level of detail in their data.

1 A retweet is a re-posting of a Tweet. It is usually identified by "RT" at the beginning of the post and is marked by Twitter's retweet icon and the name of the user who retweeted.

2 An API is a platform provided by the social networks allowing other applications and websites to pull the social media data and integrate with their site or application.

The Development of the Three-Phase Methodology

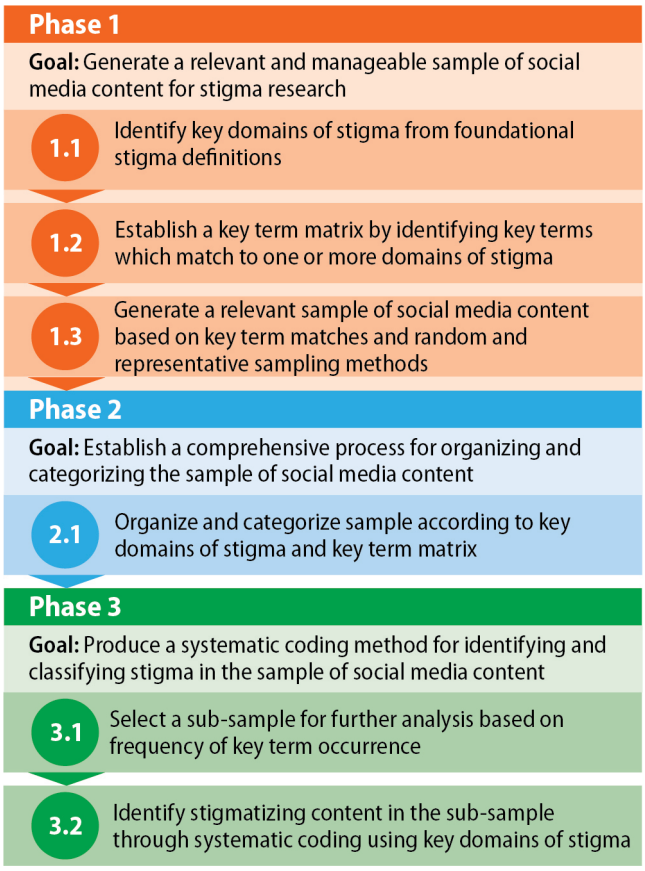
This research applied existing best practices in social media and stigma research and stigma reduction interventions, including where the two overlapped. Before beginning work, RTI's Institutional Review Board made a not-human-research determination for this research. The proposed three-part methodology uses core definitions of stigma to (1) define a relevant and manageable sample of content from social media for stigma identification and research; (2) establish a categorization process to organize the sample; and (3) guide a systematic coding method for classifying stigma within the sample (Figure 1). The following sections will describe the development of each phase of the process and the steps that define each phase.

Phase 1: Generate a Relevant and Manageable Sample of Social Media Content for Stigma Research

One of the primary challenges with social media–related research is that the sheer volume of data generated on social media platforms makes it difficult to select samples that are both representative of the diversity in the data and still relevant to the particular research focus. Existing strategies for combatting stigma focus foremost on successfully identifying and naming stigma as a way to better understand and address the key drivers of that stigma (Asadi-Aliabadi et al., 2020; Huda et al., 2020; Krishnatray, 2020; Logie, 2020; WHO et al., 2020; Yuan et al., 2021). However, current research on COVID-19 stigma on social media is sparse; the studies that do exist take a limited approach to what content is considered stigmatizing and tend to focus on occurrences of only pre-selected stigmatizing language (Budhwani

 https://doi.org/10.3768/rtipress.2023.op.0087.2305

Figure 1. Step-by-step methodology for sampling, identifying, and classifying COVID-19 stigma on social media

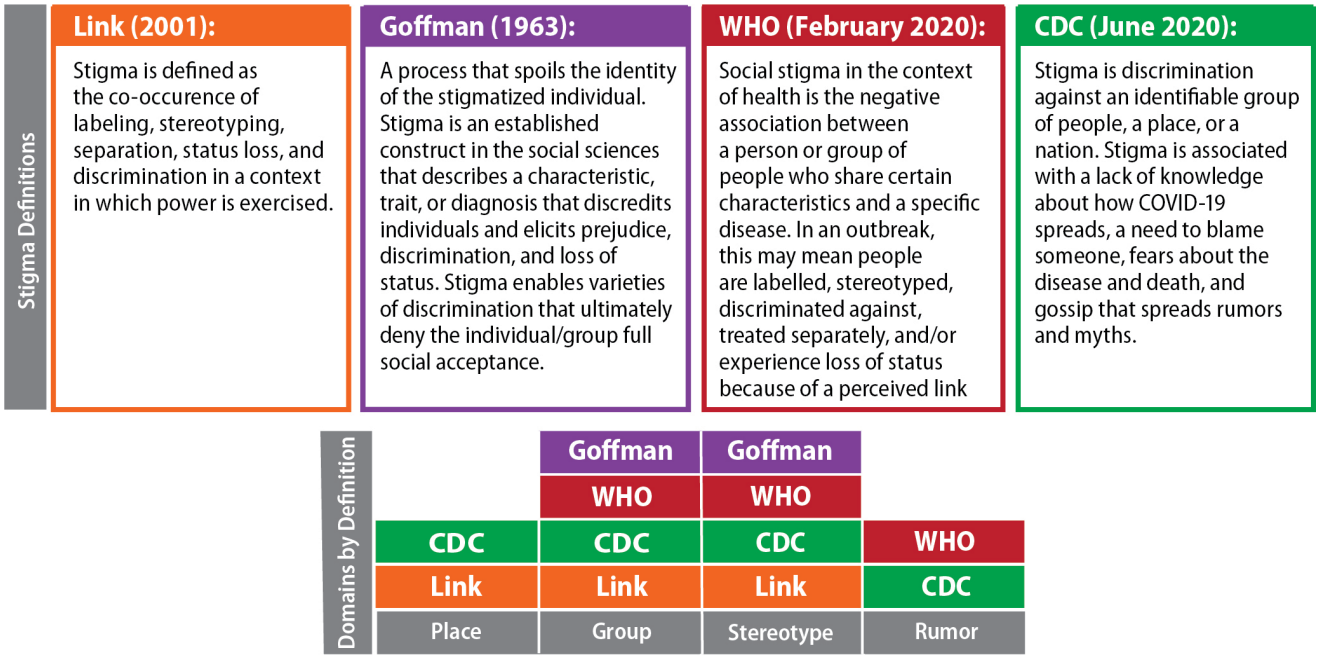


& Sun, 2020; Huda et al., 2020; Li et al., 2020) rather than the larger issue of identifying and naming COVID-19 stigma on social media in multiple forms and manifestations. To address this gap in current COVID-19 stigma on social media research, the goal of Phase 1 of this methodology was to use evidence-based definitions of stigma and a corresponding framework to generate a social media sample that is relevant and applicable to stigma research. Given the anticipated total available data points (hundreds of millions), an additional goal of Phase 1 was to develop a replicable sampling process to arrive at a more manageable sample size.

Step 1.1: Identify Key Domains of Stigma From Foundational Stigma Definitions

We conducted a scoping literature review to understand social media research best practices, stigma identification and classification methods, and where they overlapped. From the review of gray literature and peer reviewed papers, we established four core and foundational stigma definitions. We then used the four core stigma definitions to identify four key domains of stigma—Place, Group, Stereotype, and Rumor (Figure 2)—and we posited that the key domains could successfully be used to

Figure 2. Four core stigma definitions as parameters for stigma domains



 https://doi.org/10.3768/rtipress.2023.op.0087.2305

identify language that suggests the presence of stigma within social media content.

Step 1.2: Establish a Key Term Matrix by Identifying Key Terms That Match to One or More Domains of Stigma

We applied the key domains of stigma as criteria to refine stigmatizing language into a set of key terms using a sample of content from Reddit and Twitter. We used Brandwatch, a commercially available social media listening tool, as the initial refining tool because it systematically reduces the number of possible datapoints through random sampling; in this research, we used Brandwatch's COVID-19 dashboard, which they created and offered to Brandwatch users in 2020. The Brandwatch COVID-19 dashboard results sampled 1.0964 percent of all available content within our selected parameters. The dashboard is built from a query, which is a collection of key terms, phrases, and hashtags.[3] The terms used for this query included "coronavirus," "#Covid19," and numerous other words and phrases relating to the COVID-19 pandemic, both in hashtag format and as regular text. The Brandwatch COVID-19 dashboard compiles posts that contain "mentions"[4] of COVID-19 across various social media platforms and forums. To generate our sample, we applied a date range to Brandwatch's COVID-19 dashboard. The final sample was limited to public posts in English from Twitter and Reddit between February 1, 2020, and February 1, 2021. This timeframe was selected because COVID-19 was first identified in Wuhan, China on December 31, 2019, and had spread to 24 countries by February 1, 2020. This 1-year time period beginning February 1, 2020, reflects when COVID-19 became more familiar and more openly discussed by the general public and ensured that posts were representative of current affairs and trending topics and to produce as many examples as possible.

Applying the additional selection criteria to the Brandwatch COVID-19 dashboard query resulted in a sample of 19.41 million Twitter and Reddit posts, which is 1.0946 percent of the total possible data points that met the query criteria (estimated at 1,770,138,636). Brandwatch's built-in data visualization features, such as a word cloud and word frequency and occurrence graphs, organized the results for further examination. The word cloud tool collates frequently used words and depicts them in different sizes, with higher occurring words appearing larger. Review of the data visualization tools and subsequent conversion to lists allowed us to identify words or phrases that (1) were most commonly used or repeated and their variations and also (2) satisfied at least one of the four key domains of stigma. For example, the term "Chinavirus" references "China," meeting the key stigma domain of Place. We identified 31 words or phrases meeting the criteria that consistently repeated throughout the content as key terms to form the preliminary key term matrix (Table 1).

Step 1.3: Generate a Relevant Sample of Social Media Content Based on Key Term Matches and Random and Representative Sampling Methods

The key term matrix was applied to the first sample of content produced by the Brandwatch COVID-19 dashboard, which refined the initial 19.41 million posts to a second sample size of 6,376,856 posts. Of the 6,376,856 possible posts, a final sample size of 150,000 posts (2.4 percent) was selected. We chose this number based on the sample size justifications and standards established by existing COVID-19 social media research (Budhwani & Sun, 2020; Li et al., 2020), and it was in line with Brandwatch export limitations. To reduce bias and ensure that this sample was random and representative of content from across the entire sampling timeline of 365 days, results were broken into segments of 12 days, and 5,000 posts were randomly selected from each segment (365 days/12 days per segment = ~30 segments * 5,000 posts per segment = 150,000 posts). Both platforms automatically assign each post a unique "post ID," or a unique code for post identification.

The results included in the Brandwatch outputs are limited to details associated with that single post, such as text, images, and timestamp. To obtain

3 On certain platforms such as Twitter, adding a "#" to the beginning of an unbroken word or phrase creates a hashtag. When you use a hashtag in a Tweet, it becomes linked to all of the other Tweets that include it.

4 This research was structured around posts which are single units of content specific to each social media site (one Tweet = one post). A "mention" is Brandwatch's term for a singular social media post; our research utilizes Brandwatch's analyses of "mentions" but refers to the same content as "posts" throughout this work.

 https://doi.org/10.3768/rtipress.2023.op.0087.2305

Table 1. Key terms by stigma domain for sample selection and organization

DOMAIN			
Place	**Group**	**Stereotype**	**Rumor**
China	Chinese	Bat	Micro-chip
Asia	Asian	Sanitation	Laboratory
Wuhan	Foreign	Anti-mask	Food
Wuhanvirus	Immigrant	Anti-vax	Hoax
Wuflu	Oriental	Pro-mask	Experiment
Chinavirus	—	Virus denier	Rush
China flu	—	Kung flu	DNA
CCP	—	Market	Communist
CCPVirus	—	Safe	—

additional information, such as popularity, reactions, retweets, and other key information, it was necessary to collect the raw data underlying the Brandwatch output from the Twitter and Reddit APIs. Although Brandwatch collates and presents the content from the social media sites concisely, it is not the original source of the data, and the API captures key metrics that are not included in Brandwatch's interface or data access. This step is called an API query or call to the API. Querying the APIs using a sample that was already screened by the Brandwatch search ensured that our sample only included results relevant to COVID-19 without additional "social media noise" (random and irrelevant results). It also allowed for a sample of content from an entire calendar year without producing an unmanageable volume of data.

The post IDs for all 150,000 posts were exported from Brandwatch's COVID-19 dashboard and were built into the API query to pull additional relevant information, including post popularity metrics, user popularity metrics, and any links or hashtags shared within the post text, if applicable. The API query of 150,000 post IDs returned 138,998 valid posts (952 Reddit posts + 6,555 Reddit comments + 131,491 tweets). The final sample size of 138,998 was lower than the original Brandwatch sample because posts could have been deleted between the time of Brandwatch's sampling and the API call or users changed their profiles to private and therefore their data became unavailable. Figure 3 summarizes the sampling process.

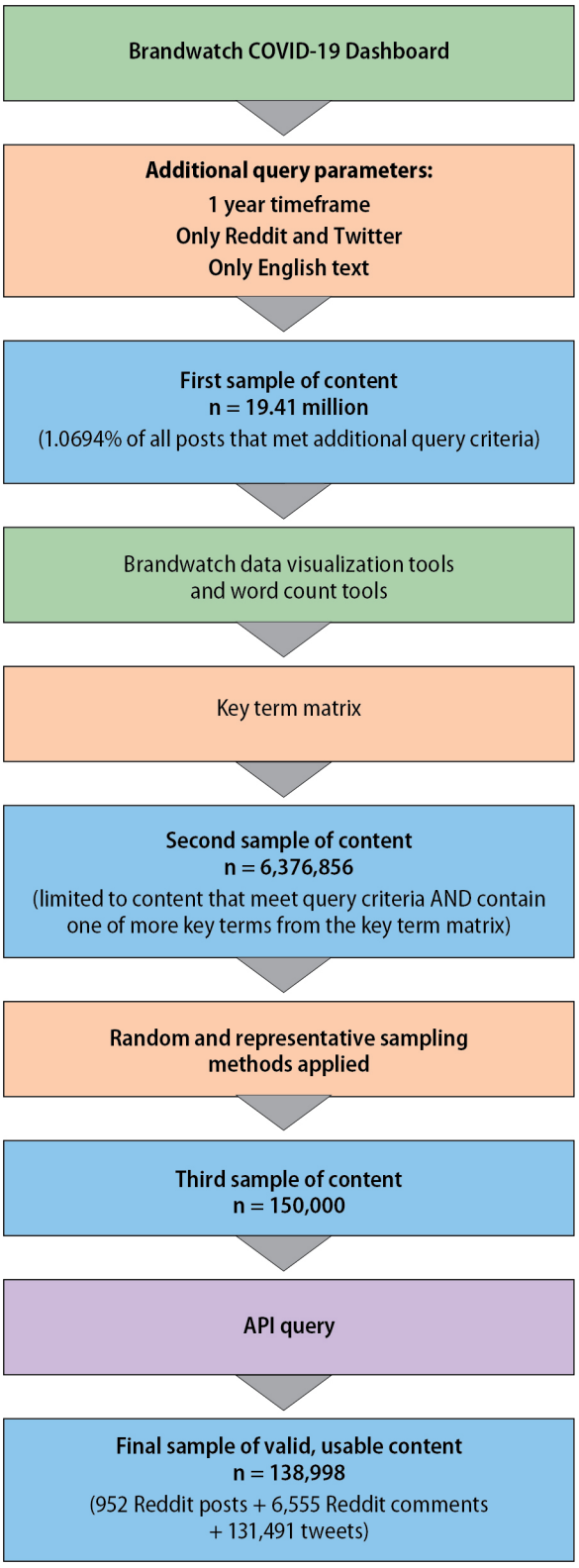
Phase 2: Establish a Comprehensive Process for Organizing and Categorizing the Sample of Social Media Content

Phase 1 of the methodology established a sample of social media content that was relevant to COVID-19 stigma and that contained a manageable number of representative data points throughout a yearlong timeframe. Phase 2 of the methodology offers comprehensive ways to organize and categorize a sample to facilitate its use for further stigma-related analyses. In this case, Phase 2 results in a process to effectively manage the resulting sample of 138,993 posts and to understand how and where stigma occurs within the content.

Step 2.1: Organize and Categorize Sample According to Key Domains of Stigma and Key Term Matrix

Using both Brandwatch and the platforms' APIs, we refined the initial potential sample of over 6 million posts to 138,993 posts from the API (Figure 3). We organized and sorted all posts by the key terms in Table 1 and the four key stigma domains in Figure 2. First, the text of the entire sample was cleaned of non-English characters, hashtags were converted to regular words, and words were unconjugated (when possible) to enable the most phrases to be found. Key term occurrence included any variations of the term, such as spelling variations, plurals, and hyphens (e.g., "anti-vax" also included "antivax," "anti-vaxxer," "antivaxx," "antivaxxers"). Because the initial query for Brandwatch's COVID-19 dashboard included

Figure 3. Process for generating a relevant and manageable sample that is both random and representative



"coronavirus," "COVID-19," and their derivatives, those terms were not included as key terms or in the sorting process.

Sorting the sample by key term and domain made it possible to understand the frequency of key term occurrence and co-occurrence of terms, therefore establishing the true relevance of the selected key terms to the sample of content. Of the 138,998 posts in the sample, Twitter posts made up 95 percent of the final sample; this skewing of the sample can be explained by the larger volume of Twitter content compared with Reddit content and because the Brandwatch COVID-19 dashboard query included hashtags, which are rare on Reddit.

Example Application of Phase 2 Methods

The results of the organization and categorization processes developed in Phase 2 underwent select descriptive analyses and are presented in Annex A. These descriptive analyses were an important step in the development of the methodology because their success was an indicator for potential replicability and feasibility of the methodology. Further refinement of the methodology could include the need to identify additional key terms or increase the sample size. The results were similar across both Twitter and Reddit, supporting applicability across different sources. Although not exhaustive, the selected analyses demonstrated success with using the methodology to refine an initially very large sample (millions of posts) into a relevant, manageable sample that can then be functionally analyzed.

Analyses relevant to the Phase 2 application (Annex A) are limited to the specific search criteria used to generate this sample but may offer an insightful example of what researchers can expect to see when studying stigmatizing social media content. For example, our results suggest that users of the methodology may expect to find that approximately 13.5 percent of posts include terms from any one of the four key stigma domains. This knowledge may aid researchers when there is a target number of posts needed that include stigmatizing content. Researchers can apply the 13.5 percent to arrive at a total sample size (if Y is the target number of needed posts potentially containing stigma, a sample size of Y/13.5 percent is best).

https://doi.org/10.3768/rtipress.2023.op.0087.2305

The frequency of the use of certain key terms and specific domains demonstrated the types of content and themes that were most prevalent within the sample. Across both platforms, 14,521 posts (10 percent) included one or more key terms, of which 83 percent (12,068 posts) included only one key term. The number and distribution of posts that included key terms from each of the four stigma domains produced similar results on both Reddit and Twitter, with Place as the domain with the highest post count (n = 7,274, or 5.5 percent of the total sample). This may suggest that, specific to COVID-19 stigma, issues surrounding xenophobia are prevalent.

An additional sub-analysis explored the use of key terms within Twitter hashtags as a potential way to organize the sample. The key terms most commonly used as hashtags were "ccp" (n = 138), "wuhanvirus" (n = 114), and "chinavirus" (n = 49), which was in line with frequently used key terms without a hashtag. Researchers initially posited that hashtags would assist in identifying additional key terms, but results showed that any difference in terminology was not significant. Because hashtags link multiple tweets and are typically used to identify trending themes, there is potential to capitalize on their unique properties to define the most current stigmatizing messaging or content. Our ability to click through numerous posts that all used the "wuhanvirus" hashtag aided the key term generation process in Phase 1, but future research would benefit from exploring the additional capabilities of using hashtags specifically to better understand stigmatizing content.

Phase 3: Produce a Systematic Coding Method for Identifying and Classifying Stigma in the Sample of Social Media Content

Phases 1 and 2 offered a manageable and relevant sample of social media content organized by the key term matrix and domains of stigma. These previous phases and their steps helped organize the sample in ways that facilitates identification and analysis of any stigma within the content. An important component of the methodology involves accounting for the potential subjectivity that can occur when researchers attempt to identify stigma. Phase 3 of this methodology aimed to test the applicability of the processes developed to identify stigmatizing content. We adapted the key domains and key terms for use as a qualitative coding tool to identify and classify occurrences of stigma within the API sample. This coding method is an important tool for systematic stigma identification because it helps standardize the process of stigma identification across different researchers.

Step 3.1: Select a Sub-sample for Further Analysis Based on Frequency of Key Term Occurrence

We sorted the API sample of 138,998 posts by posts with the highest occurrence of key terms as part of the Phase 2 organization and categorization methods. Because the key terms were selected based on their relevance to the four key domains of stigma, we hypothesized that selecting posts with the highest occurrences of key terms would increase the likelihood that those posts would contain stigmatizing content. Of the original API sample of 138,998, 711 posts had two or more key term occurrences, across any combination of domains; we selected these 711 posts as the sub-sample for the content analysis. Although some of the key terms in themselves could be classified as stigmatizing, a higher key term occurrence would also avoid counts of key terms that are not inherently stigmatizing as standalone terms (e.g., "market" or "China").

Step 3.2: Identify Stigmatizing Content in the Sub-sample Through Systematic Coding Using Key Domains of Stigma

We [NW, PM, MM] created a preliminary codebook with codes and definitions for each of the four key domains of stigma, one code for no stigma, and two codes for stigma intentionality (Table 2). We included the stigma intentionality codes to account for the constantly emerging and changing understanding of language surrounding COVID-19. For example, early in the pandemic the term "Wuhanvirus" was used by news sources and even academics (Su et al., 2020) before it was more widely recognized as a stigmatizing term. The use of this term required a code that recognized it as stigmatizing but also acknowledged the likely intention of its use. We discussed the parameters on how to define the stigma intentionality codes and worked to agree on different scenarios.

Inclusion of a code for no stigma was important due to the limitations of the methods established in

Table 2. Codes for identifying and classifying stigma within the sample of social media content

Code	Definition
Intentional Stigma	*Use this code in combination with a domain code for text that intentionally seeks to stigmatize.*
Unintentional Stigma	*Use this code with a domain code when the text is stigmatizing, but it is not intentional in nature (i.e., using "wuhan coronavirus" before there was a name for COVID-19).*
Group	[DOMAIN] *Use this code when the text includes a reference to a group of people, including references to people who are foreign, immigrants, Asian, and Chinese.*
Place	[DOMAIN] *Use this code when the text includes a reference to place, including references to China or Wuhan.*
Stereotype	[DOMAIN] *Use this code when the text includes a reference to a stereotype, or a generalization of a group of people. This could include mentions of pro-mask, anti-vaxxer, unsanitary, or safe.*
Rumor	[DOMAIN] *Use this code when the text includes a reference to a COVID-related rumor or misinformation, such as micro-chip, laboratory, or hoax.*
No stigma	*Use this code when the text does not include a reference to the stigma domains.*

Phase 2. The use of the key term matrix as parameters for sample selection helped ensure that all posts in the sample would be relevant to COVID-19 and contained one or more key terms that could *potentially* be associated with stigma. This association was important for narrowing a very large sample, but the presence of any key terms within the content does not guarantee that stigma was present. In these cases, the no stigma code classified content that met the initial search criteria because it included a term from the key term matrix but was ultimately non-stigmatizing. For example, the key term matrix included the term "China," which appeared often in the posts that contained stigma but also occurred frequently in other types of content, such as news articles or factual information, without any stigmatizing elements.

Three of the researchers [NW, PM, MM] randomly selected 20 posts from the sample of 711 to independently code using Atlas.ti and then assessed inter-coder agreement through reviewing each person's coding of the 20 posts. After discussion and comparison, we refined the codebook and then tested on an additional random 10 posts. Following this final inter-coder exercise, the total sample of 711 posts was randomly divided for each researcher to individually code 237 posts to result in a final codebook.

We designed the codebook to align with the sample selection and organization processes by matching the content in the posts to specific elements of the definitions of stigma (i.e., the domains). By relying on standard definitions and domains of stigma, researchers could independently and efficiently code the selection of 711 posts (237 posts each). This step ensured that when content was coded as stigma (intentional or unintentional), selection of a domain code clarified what components of the content were considered stigmatizing and how. This coding method was driven by the stigma definitions and domains established in Phases 1 and 2 and supports identifying stigmatizing content in a standardized and replicable approach between researchers.

Example Application of Phase 3 Methods

Here, and in Annex B, we present examples of the application of Phase 3 of the methodology with analyses on how stigma manifested within the sample of social media content. When Twitter and Reddit users post content with public privacy settings, user policy suggests that they understand and accept the public nature of their content. However, to further protect the creators of the content we have sampled, we have anonymized and paraphrased the posts so they are not an exact text match to the originals. Our analysis of the 711 posts from Reddit and Twitter yielded notable examples of stigmatizing content. Analyzing the sample through a content analysis of the entire post beyond the text to include reach, timeline, and other platform characteristics allowed us to better understand the context in which stigmatizing language may be used. The ability to systematically confirm the existence of stigma through coding with specific and concrete stigma domains facilitates more objective studying of stigma on social media.

 https://doi.org/10.3768/rtipress.2023.op.0087.2305

Determining Stigma Intentionality. Intentional stigma addresses the assumed motive of the content creator. In our analyses, coding a post as intentional stigma was determined by its use of any of the key domains of stigma according to prominent definitions of stigma and was related to current knowledge about appropriate language and terms related to COVID-19 at the time of the post. Intentional and unintentional stigma can shift depending on current knowledge and standard practices. For example, before February 2020, the WHO had yet to announce the formal name as COVID-19 and so use of "Wuhan Virus" or "China Virus" were more common. COVID-19 was later chosen as the official name to intentionally avoid using names that referred to place, group, or people, so intentional use of these terms after February 2020 can theoretically be determined as potentially intentional stigma. An example of this can be seen in the tweet in Figure 4, particularly the hashtags, which was dated from March 18, 2020.

We also classified content as intentional stigma when it included language that was purposely crafted to be stigmatizing through the use of one or more of the direct domains of stigma. Examples include racist nomenclature to refer to COVID-19, such as a post including the phrase "kung flu-k you, China!," which includes the stigma domain of Place with the suggestion of blame.

We coded this as intentional stigma because it uses an unofficial term for COVID-19 deemed racist ("kung flu") and established negative sentiment and blame associated with Place (China) (word choice or play on words using offensive terms).

According to the codebook, researchers were instructed to use the intentionality codes based on the content, not any accompanying user or post information. However, while analyzing text for the use of the key terms and how they were applied, researchers also inadvertently picked up on tone, mood, or intended audience, and this could have influenced the use of the intentionality codes. This limitation of human coding may have contributed to how we interpretated or perceived the content. For example, we coded this tweet as unintentional stigma and with the domain codes of Place and Rumor:

Figure 4. Example of potentially intentional stigma

> I'm on the same page with Hong Kongers to end the fascist #CCP. I made it through Mao's Cultural Revolution as an adolescent and then found freedom in America. I thought I was safe now. I was incorrect. #Xitler released #WuhanVirus to kill 136,000 US citizens in a bio-attack. I might be the next person. We have no choice but to fight.

This post meets our definition of stigma through the use of terms that fit into the domains of Place and Stereotype such as "WuhanVirus" and "Xitler" (a play on the names of Adolf Hitler and President Xi of China). The tweet also meets the criteria for the Rumor domain, asserting that COVID-19 was manufactured and released to purposefully kill Americans. This tweet was among a few examples where the purpose of the post was political ideation or activism. The user self-identifies as a Hong Kong native residing in the United States and has expressed negative sentiment regarding the Chinese Communist Party (CCP) and Former President Mao. Coding of this post followed protocol and was coded as stigmatizing because of the presence of the domains. However, the coder also understood the post's purpose as likely political and therefore applied the unintentional stigma code. This distinction is important because it (1) recognizes the intentionality of using terms to imply rumor and blame toward the CCP and President Mao, and (2) by using stigmatizing terms at all, this post contributes to overall stigma toward Chinese people or Asian people, even if this was not the author's intent. The references to communism, fascism, and

bio-attacks may be less about the people of China or Asian people and instead be more specifically about CCP as a political institution. This example demonstrates the value of the systematic coding system for identification of stigma; had we not completed a content analysis, this post would have met criteria for stigmatizing language without a further understanding of why. It also presents another consideration for limitations of the methodology, including the gaps in and importance of deciding how to incorporate context when coding.

Naming the Elements of Stigma Using Stigma Domains. The content analysis allowed for detailed exploration of how stigma occurred within the sample. Using the key stigma domains as the primary codes for the content analysis ensured that evidence-based stigma definitions drove the identification of stigma within the sample. During the coding process, we determined that content inclusive of only one domain of stigma, even with multiple key terms, was more difficult to determine as stigmatizing and as either unintentional or intentional stigma. As a result, we recommend that criteria for the systematic coding method should require two of more domains of stigma to be considered a strong match for stigmatizing content. The Reddit comment in Figure 5 is a good example of the application of multiple domains.

This comment from a Reddit post includes reference to Place (China) in direct connection with blame (directed at China, at "Progressives," Former President Barack Obama). It also includes the domains of Stereotype and Rumor in the assertations about laboratories and laboratory safety and the malicious creation of COVID-19.

Additionally, some examples demonstrated that the co-occurrence of multiple domains of stigma is what made content stigmatizing. We provide an example of this in the Reddit comment in Figure 6.

We coded this example as stigmatizing because of the relationship between the domains. In the previous example, the use of Place: "Why not hold China responsible?" is stigmatizing on its own, as is the use of Rumor: "They created the virus in a laboratory." However, in the second example,

Figure 5. Example with stigma across multiple domains

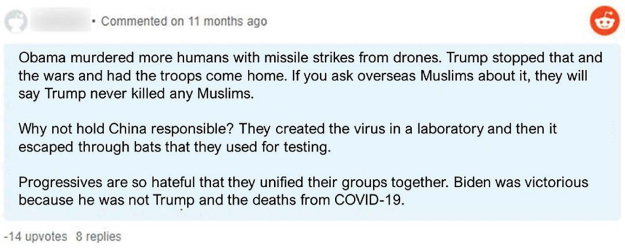
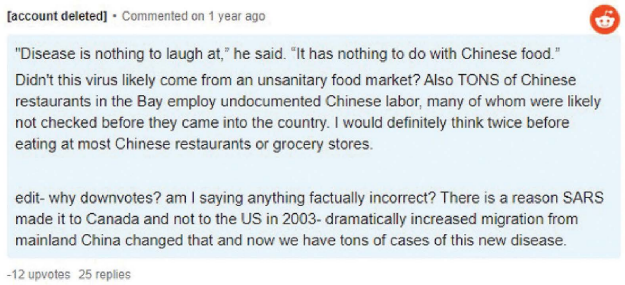


Figure 6. Example of co-occurrence of multiple domains creating stigma



the content refers to Chinese restaurants, people, and employees—the Group domain. It also meets the stigma domains of Rumor and Stereotype by implying these groups are responsible for spreading COVID-19. It is the use of these domains in connection with one another, not just sheer mention of the groups, which make the content intentionally stigmatizing. Mentioning just Group or Rumor and Stereotype individually may not independently qualify as stigma but used in combination and with the intention of blame, the content meets full stigma criteria. Phase 3 of the methodology recognizes that drivers of stigma are often complex and overlapping and identifying stigma requires a process to both name each of the elements of stigmatizing content and the impact of the coalescence of those elements.

Discussion

Stigma is a complex concept caused by a myriad of factors and presents itself in diverse ways with various outcomes. Infectious disease stigma specifically can affect health-seeking behaviors, availability of services, access to services, and many other consequential outcomes (Nyblade et al., 2019). Identifying stigmatizing social media content is even more

complex because of the natural subjectivity of what users post. Recent work regarding COVID-19 stigma and social media has emerged but has lacked replicable systematic methods for identifying and categorizing diverse presentations of stigma (Budhwani & Sun, 2020; Di et al., 2021). A single definition, methodology, or framework that systematically identifies social media content as stigmatizing is a necessary step to allow for identification of points of intervention. Understandably, there are many barriers to developing a single tool for stigma identification that relate to its inherent complexity and subjectivity. With the development of this methodology, we aimed to address some of these challenges with methods for managing the sheer volume of available data points and a process to identify and classify stigmatizing content.

Significance of Findings

This research resulted in a proposed methodology that offers guidance to future researchers on: (1) how to generate a relevant and manageable social media social sample for stigma identification and research, (2) a categorization process to organize the sample, and (3) a systematic coding method for classifying stigma within the sample.

The distribution of posts that contained key terms from each of the four key stigma domains (Place, Group, Stereotype, Rumor) was similar across both Reddit and Twitter. This significant finding demonstrates the methodology's success across two social media platforms that differ in many ways, including approach, engagement, organization, and value to the user. Restrictions to API and data access aside, this finding suggests that the methodology may be applicable to additional platforms.

Of the four key stigma domains, Place had the highest associated post count (Annex A). We created the Place domain to reflect the distinct elements of stigma definitions that include labeling, separating, and discriminating against people because of their association with a specific place or nation. The high occurrence of Place is likely a reflection of the overall sentiment and attention focused on China and Asia as the verified origin of the disease. The resulting stigma and discrimination against Asian Americans and people of Asian descent has been pervasive throughout the COVID-19 pandemic. This type of xenophobia is closely connected to the type of stigma that the key stigma domain of Place often indicates. The cyclical nature of the relationship between Place and stigma creates difficulty in understanding which occurs first—does xenophobia lead to stigmatizing behaviors, or does the stigma that exists as a result of other factors increase the potential for xenophobia?

The high prevalence of Place, represented by references to China or other Asian countries, in the sample may also reflect the presence of intersectional stigma, or when a convergence of multiple stigmatized identities within a person or group exists (Turan et al., 2019). Although this work did not fully analyze the prevalence and presence of intersectional stigma within our sample, it is an important component of understanding stigmatizing content. The content analysis revealed COVID-19–related stigma in connection with Place, including many examples of stigma related to the socio-political environment in China and tense relationships between "Western powers" and China or the CCP. Such content addressed multiple overlapping identities of particular individuals or groups, such as nationality, political ideology, and racial identity. References to multiple stigmatized identities suggest the presence of intersectional stigma. Further analysis would be useful to better understand how the proposed identification and coding methods can facilitate the recognition of intersectional stigma within social media content.

The importance of addressing unintentional stigma is demonstrated in evidence-based examples of stigma reduction approaches; awareness and attitudes are well-recognized as key drivers of stigma (Earnshaw & Chaudoir, 2009; Nyblade et al., 2021; Nyblade et al., 2019; Stangl et al., 2019). Interventions that address people's awareness and attitudes can help them realize ways in which they were perpetuating stigma without even realizing it (Kim et al., 2018; Nyblade et al., 2019; Stangl et al., 2019). Existing research clearly demonstrates the impact of unintentional stigma (Nyblade et al., 2019), and the stigma intentionality codes in Phase 3 address the concept of intent as a requirement to identify stigma or not, but further research that specifically highlights

stigma intentionality in social media content would strengthen this approach. Given that 67 percent of the stigmatizing posts were coded as unintentional, we stress the relevance of understanding intent within stigmatizing social media content.

Misinformation is also highly prevalent on social media, particularly surrounding health issues, and has the power to shape people's perceptions and understanding of key issues, without much fact checking or adherence to truth (Suarez-Lledo & Alvarez-Galvez, 2021). We included the stigma domains Stereotype and Rumor as codes to account for the presence of misinformation and disinformation and their unique relationship to stigmatizing content. Based on the core definitions of stigma, misinformation and lack of awareness are fundamental contributors to stigma (Nyblade et al., 2019). This idea is reinforced by the fact that interventions that focus on empowering people with knowledge and accurate information continue to be the most effective way to combat stigma (Earnshaw & Chaudoir, 2009; Nyblade et al., 2021; Nyblade et al., 2019; Stangl et al., 2019). A compelling reason to advocate for further research and refinement of methodologies like the one proposed in this paper is the idea that generating comprehensive processes for addressing stigma on social media may also help address the problem of misinformation on social media. Additionally, efforts by researchers, policy makers, and even the governing bodies and management of social media platforms may find that efforts to address misinformation on social media would benefit from the best practices that have been established in stigma research.

Scope for Future Use and Scale-up

The process to develop and apply this methodology offers a useful starting point for future COVID-19 stigma research on social media and could potentially be applicable to other types of stigmas as well. The establishment of stigma domains and then translation of those domains into a key term matrix and a codebook provides a more systematic way to identify stigma across large samples of social media content and potentially other forms such as print media.

Further research that examines broader health-related stigma, such as mental health, HIV, AIDS, and contraception or abortion care, may apply the same domains to generate more relevant key terms in social media content. Replication or application of the methodology for other types of stigmas should begin with further exploration of existing literature or research specific to that type of stigma because the domains or key terms would likely differ from the ones used in this methodology. Depending on whether intersectional stigma may exist, future research that uses this methodology would benefit from developing key terms that address other traits or identities. Specific to COVID-19 as an infectious disease, for example, stigmatizing content could include specific references to the elderly, homeless or housing insecure populations, relevant professionals, or other potentially stigmatized groups.

This methodology offers a comprehensive way to generate relevant social media content samples and as demonstrated, it identified 13.5 percent of the total sample as containing one or more key terms. However, even being able to narrow the sample down to this extent will leave researchers with many data points to examine because of the sheer volume of social media content. Integrating methodologies like this one into the management of social media data may be essential to sorting data at this scale. Recently, social media platforms like Facebook and Twitter have been under fire for the lack of content moderation on their sites, because of either an inability or unwillingness to do so (Bond, 2021). There remain many outstanding questions about the regulations and policies of online environments, accountability of content, and how researchers and policymakers can use the methodology to hold social media platforms accountable.

These types of sorting and classification processes are increasingly important in social media research because the scope and breadth of the potential content is far larger than almost any other type of language-based data source. When the potential sample sizes for social media research are rarely smaller than millions of posts, initial sorting techniques and strategies become vital to ensuring that samples are manageable, representative, and

relevant to the focus of the research. Social media will only continue to be ever-present in society, thus the potential for use of the proposed methodology to address stigma as one of the most pervasive barriers to healthcare is vast and timely.

Limitations

This methodology is an important contribution to the inadequate research on stigma and social media, and additional research would address many of the limitations. Several limitations of the methodology are mentioned in connection with the analysis provided previously. More concrete limitations include that our development and application of the methodology was limited to Twitter and Reddit, whose users may not be representative of the general population, and we only included an initial sample of 150,000. Expansion to other platforms or increasing the sample size may assist in determining additional key terms that fall into a domain or contribute to stigmatizing language on social media. The application of the methodology also included results across an entire year (February to February) with the intention of capturing as wide a variety of content as possible but ultimately producing an extremely large number of posts. Selecting a shorter timeframe could increase the number of posts that are identified as stigmatizing.

We decided to limit the sample to English language content because the tools available to use were designed to work best with English content but also because there was ample content available in English to analyze from Twitter and Reddit. The decision to only use English content had a significant impact on development of the key term matrix, which heavily influenced results to reflect anti-Asian stigma that has been present predominantly in the Western context. We note that this limitation stresses the importance of the step to develop the key term matrix, including the consideration of what languages to include and how language might impact results.

As with any content analysis requiring contextualization, a clear limitation of this work is that we did not apply coder cross-examination in Phase 3 to the whole sample. The coders attempted to systematically code for the elements of stigma and the authors' intention in each post using the established and agreed upon codebook, but use of the stigma intentionality codes leaves room for differing interpretations and judgements of each individual coder. We collectively assessed inter-coder reliability and discussed a subset of the content and use of codes, but we did not quantify a metric of inter-coder reliability. After adjusting the codebook following the initial assessment, we only re-coded and qualitatively reassessed a few sources that were co-coded. We view this limitation of the qualitative analysis as easily addressable in future research, largely because the purpose of the qualitative analysis completed for this paper was process-oriented. If future methods were adapted to focus on the results, then more robust inter-coder agreement would be needed.

Each phase of the methodology serves a specific purpose as part of the overall methodology goals and does not produce the same or complete results if used independently of each other. Put another way, the methodology is unlikely to be successful if both the creation of domains and the key term matrix are not included. Similarly, development of the key term matrix without the quantitative analysis and qualitative assessment techniques used for coding is unlikely to outright determine the existence of stigmatizing content.

Conclusion

Social media and stigma research—combined and independently—are growing bodies of work, and tools, methods, or frameworks that support strong research in these areas are vital. Our experiences as social media users at the onset of the COVID-19 pandemic made it very clear that COVID-19–related content contained negative messaging and stigmatizing language and must be addressed. The identification of stigmatizing content among potentially millions of posts is crucial to the success of any response intervention, and this methodology aims to help researchers find evidence-based solutions using a manageable sample.

References

Asadi-Aliabadi, M., Tehrani-Banihashemi, A., & Moradi-Lakeh, M. (2020). Stigma in COVID-19: A barrier to seek medical care and family support. *Medical Journal of the Islamic Republic of Iran*, *34*, 98. https://doi.org/10.47176/mjiri.34.98

Azim, D., Kumar, S., Nasim, S., Arif, T. B., & Nanjiani, D. (2020). COVID-19 as a psychological contagion: A new Pandora's box to close? *Infection Control and Hospital Epidemiology*, *41*(8), 989–990. https://doi.org/10.1017/ice.2020.127

Bloomberg. (2020). *Social stigma and harassment undermine COVID-19 testing efforts across Asia*. The Japan Times. https://www.japantimes.co.jp/news/2020/05/13/asia-pacific/stigma-harassment-coronavirus-testing-asia/

Bond, S. (2021). *Facebook, Twitter, Google CEOs testify before Congress: 4 things to know*. National Public Radio. https://www.npr.org/2021/03/25/980510388/facebook-twitter-google-ceos-testify-before-congress-4-things-to-know?t=1657811262456

Budhwani, H., & Sun, R. (2020). Creating COVID-19 stigma by referencing the novel coronavirus as the "Chinese virus" on Twitter: Quantitative analysis of social media data. *Journal of Medical Internet Research*, *22*(5), e19301. https://doi.org/10.2196/19301

Churcher, S. (2013). Stigma related to HIV and AIDS as a barrier to accessing health care in Thailand: A review of recent literature. *WHO South-East Asia Journal of Public Health*, *2*(1), 12–22. https://doi.org/10.4103/2224-3151.115829

Di, Y., Li, A., Li, H., Wu, P., Yang, S., Zhu, M., Zhu, T., & Liu, X. (2021). Stigma toward Wuhan people during the COVID-19 epidemic: An exploratory study based on social media. *BMC Public Health*, *21*(1), 1958. https://doi.org/10.1186/s12889-021-12001-2

DuPont-Reyes, M. J., Villatoro, A. P., Phelan, J. C., Painter, K., & Link, B. G. (2020). Media language preferences and mental illness stigma among Latinx adolescents. *Social Psychiatry and Psychiatric Epidemiology*, *55*(7), 929–939. https://doi.org/10.1007/s00127-019-01792-w

Earnshaw, V. A., & Chaudoir, S. R. (2009). From conceptualizing to measuring HIV stigma: A review of HIV stigma mechanism measures. *AIDS and Behavior*, *13*(6), 1160–1177. https://doi.org/10.1007/s10461-009-9593-3

Ekstrand, M. L., Bharat, S., & Srinivasan, K. (2018). HIV stigma is a barrier to achieving 90-90-90 in India. *The Lancet. HIV*, *5*(10), e543–e545. https://doi.org/10.1016/S2352-3018(18)30246-7

Gesesew, H. A., Tesfay Gebremedhin, A., Demissie, T. D., Kerie, M. W., Sudhakar, M., & Mwanri, L. (2017). Significant association between perceived HIV related stigma and late presentation for HIV/AIDS care in low and middle-income countries: A systematic review and meta-analysis. *PLoS One*, *12*(3), e0173928. https://doi.org/10.1371/journal.pone.0173928

Gourlay, A., Birdthistle, I., Mburu, G., Iorpenda, K., & Wringe, A. (2013). Barriers and facilitating factors to the uptake of antiretroviral drugs for prevention of mother-to-child transmission of HIV in sub-Saharan Africa: A systematic review. *Journal of the International AIDS Society*, *16*(1), 18588. https://doi.org/10.7448/IAS.16.1.18588

Hamilton, A., Shin, S., Taggart, T., Whembolua, G. L., Martin, I., Budhwani, H., & Conserve, D. (2019). HIV testing barriers and intervention strategies among men, transgender women, female sex workers and incarcerated persons in the Caribbean: A systematic review. *Sexually Transmitted Infections*, *96*(3), 189–196.

Hatzenbuehler, M. L., Phelan, J. C., & Link, B. G. (2013). Stigma as a fundamental cause of population health inequalities. *American Journal of Public Health*, *103*(5), 813–821. https://doi.org/10.2105/AJPH.2012.301069

Huda, M. N., Islam, R., Qureshi, M. O., Pillai, S., & Hossain, S. (2020). Rumour and social stigma as barriers to the prevention of coronavirus disease (COVID-19): What solutions to consider? *Global Biosecurity*, *2*. https://doi.org/10.31646/gbio.78

Jan, T. (2020, May 19). Asian American doctors and nurses are fighting racism and the coronavirus. *The Washington Post*. https://www.washingtonpost.com/business/2020/05/19/asian-american-discrimination/

Kim, H. Y., Grosso, A., Ky-Zerbo, O., Lougue, M., Stahlman, S., Samadoulougou, C., Ouedraogo, G., Kouanda, S., Liestman, B., & Baral, S. (2018). Stigma as a barrier to health care utilization among female sex workers and men who have sex with men in Burkina Faso. *Annals of Epidemiology*, *28*(1), 13–19. https://doi.org/10.1016/j.annepidem.2017.11.009

Krishnatray, P. (2020). *COVID-19 is leading to a new wave of social stigma*. The Wire. https://thewire.in/society/covid-19-social-stigma

Li, Y., Twersky, S., Ignace, K., Zhao, M., Purandare, R., Bennett-Jones, B., & Weaver, S. R. (2020). Constructing and communicating COVID-19 stigma on Twitter: A content analysis of tweets during the early stage of the COVID-19 outbreak. *International Journal of Environmental Research and Public Health*, *17*(18), 6847. https://doi.org/10.3390/ijerph17186847

Logie, C. H. (2020). Lessons learned from HIV can inform our approach to COVID-19 stigma. *Journal of the International AIDS Society*, *23*(5), e25504. https://doi.org/10.1002/jia2.25504

Nyblade, L., Mingkwan, P., & Stockton, M. A. (2021). Stigma reduction: An essential ingredient to ending AIDS by 2030. *The Lancet. HIV*, *8*(2), e106–e113. https://doi.org/10.1016/S2352-3018(20)30309-X

Nyblade, L., Stockton, M. A., Giger, K., Bond, V., Ekstrand, M. L., Lean, R. M., Mitchell, E. M. H., Nelson, R. E., Sapag, J. C., Siraprapasiri, T., Turan, J., & Wouters, E. (2019). Stigma in health facilities: Why it matters and how we can change it. *BMC Medicine*, *17*(1), 25. https://doi.org/10.1186/s12916-019-1256-2

Reddit Inc. (2022). *Redditinc.com*. https://www.redditinc.com/

Rueda, S., Mitra, S., Chen, S., Gogolishvili, D., Globerman, J., Chambers, L., Wilson, M., Logie, C. H., Shi, Q., Morassaei, S., & Rourke, S. B. (2016). Examining the associations between HIV-related stigma and health outcomes in people living with HIV/AIDS: A series of meta-analyses. *BMJ Open*, *6*(7), e011453. https://doi.org/10.1136/bmjopen-2016-011453

Stangl, A. L., Earnshaw, V. A., Logie, C. H., van Brakel, W., Simbayi, L. C., Barré, I., & Dovidio, J. F. (2019). The Health Stigma and Discrimination Framework: A global, crosscutting framework to inform research, intervention development, and policy on health-related stigmas. *BMC Medicine*, *17*(1), 31. https://doi.org/10.1186/s12916-019-1271-3

Su, Z., McDonnell, D., Ahmad, J., Cheshmehzangi, A., Li, X., Meyer, K., Cai, Y., Yang, L., & Xiang, Y.-T. (2020). Time to stop the use of "Wuhan virus," "China virus" or "Chinese virus" across the scientific community. *BMJ Global Health*, *5*(9), e003746. https://doi.org/10.1136/bmjgh-2020-003746

Suarez-Lledo, V., & Alvarez-Galvez, J. (2021). Prevalence of health misinformation on social media: Systematic review. *Journal of Medical Internet Research*, *23*(1), e17187. https://doi.org/10.2196/17187

Turan, J. M., Elafros, M. A., Logie, C. H., Banik, S., Turan, B., Crockett, K. B., Pescosolido, B., & Murray, S. M. (2019). Challenges and opportunities in examining and addressing intersectional stigma and health. *BMC Medicine*, *17*(1), 7. https://doi.org/10.1186/s12916-018-1246-9

Turan, J. M., & Nyblade, L. (2013). HIV-related stigma as a barrier to achievement of global PMTCT and maternal health goals: A review of the evidence. *AIDS and Behavior*, *17*(7), 2528–2539. https://doi.org/10.1007/s10461-013-0446-8

Twitter. (2022). *Twitter earnings report: FY2022Q2*. https://s22.q4cdn.com/826641620/files/doc_financials/2022/q2/Final_Q2'22_Earnings_Release.pdf

Wang, Y., McKee, M., Torbica, A., & Stuckler, D. (2019). Systematic literature review on the spread of health-related misinformation on social media. *Social Science & Medicine*, *240*, 112552. https://doi.org/10.1016/j.socscimed.2019.112552

World Health Organization (WHO). (2020a, February 23). *WHO situation report: Coronavirus Disease, 2019*. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200223-sitrep-34-covid-19.pdf

World Health Organization (WHO). (2020b). *WHO conducts remote psychological first aid training in Iraq to address COVID-19 stigma and discriminatory behaviours*. http://www.emro.who.int/iraq/news/who-conducts-remote-psychological-first-aid-training-in-iraq-to-address-covid-19-stigma-and-discriminatory-behaviours.html

World Health Organization (WHO), International Federation of Red Cross and Red Crescent Societies (IFRC), & United Nations Children's Fund (UNICEF). (2020). *Social stigma associated with COVID-19: A guide to preventing and addressing social stigma*. https://www.who.int/publications/m/item/a-guide-to-preventing-and-addressing-social-stigma-associated-with-covid-19

Yuan, Y., Zhao, Y.-J., Zhang, Q.-E., Zhang, L., Cheung, T., Jackson, T., Jiang, G.-Q., & Xiang, Y.-T. (2021). COVID-19-related stigma and its sociodemographic correlates: A comparative study. *Globalization and Health*, *17*(1), 54. https://doi.org/10.1186/s12992-021-00705-4

Appendix

Annex A. Select Quantitative Analyses

Summary of results

	n	%
Total posts with exactly one term	12,068	83
Total posts with more than one key term from exactly one of the four groups	822	6
Total posts with at least one key term in multiple groups	1,632	11
Total posts with any terms	14,521	10
Total Posts	138,998	100

Summary of sample from Twitter and Reddit organized by key term and domain

	Twitter		Reddit *		TOTAL
	n	%	*n*	%	*n*
Total posts	131,491	95	7,507	5	138,998
Total posts with any terms	13,957	96	564	4	14,521
Total posts with exactly one term	11,603	96	465	4	12,068
Total posts with more than one term overall	2,354	96	100	4	2,454
Total posts with more than one key term from exactly one of the four domains	794	97	28	3	822
Total posts with at least one key term in multiple domains	1,560	96	72	4	1,632

* Includes both Reddit comments and posts. The proportion of Reddit comments to posts was about 8 to 1; this occurs because a single Reddit post could contain thousands of comments.

Distribution of tweets across the number of domains

No. of Domains	Tweet count
0	120,043
1	12,397
2	1,363
3	187
4	10

Count by domain

Domain	Post count	
	n	Percentage of total sample (%)
Place	7,274	5.5
Group	2,778	2.1
Stereotype	4,323	3.3
Rumor	3,357	2.6
Total Domain Use	**17,732**	**13.5**

* Total domain use does not equal total number of posts with any key term because a single post may be duplicated across more than one domain.

https://doi.org/10.3768/rtipress.2023.op.0087.2305

Posts sorted by key term stigma domain, by platform

Domain	Twitter		Reddit	
	n	%	*n*	%
Place	7,036	41.4	238	32.6
Stereotype	4,099	24.1	224	30.6
Rumor	3,207	18.9	150	20.5
Group	2,659	15.6	119	16.3
Total Domain Use	17,001	100.0	731	100.0

Most common hashtags from sample

Hashtag	Occurrence
#coronavirus	89
#china	80
#covid19	62
#ccpvirus	42
#wuhan	35
#wuhancoronavirus	31
#coronavirusoutbreak	24
#ccpchina	20
#hongkong	12

Annex B. Select Content Analyses

Posts selected for content analysis: 711 (0.5 percent of total sample).

Total number of posts by platform

Platform	*n*	%
Twitter	610	85.7
Reddit	101	14.2
Total	711	—

Stigmatizing posts coded as intentional or unintentional stigma

Indicator	*n*	Percentage (%)
Not Stigmatizing	335	47.1
Stigmatizing	268*	37.6
Intentional Stigma	109	40.7
Unintentional Stigma	179	66.8
Total Posts Analyzed	711	—

* The total number of posts coded as stigmatizing does not equal the sum of the posts coded as either intentional or unintentional because different lines within a post were coded with both terms.

Posts coded as stigmatizing by domain

Domain	*n*	% (*n* = 711)
Place	260	36.5
Group	255	35.8
Stereotype	53	7.4
Rumor	169	23.7

Additional analyses for posts determined as stigmatizing

	n (%) of all posts (*n* = 711)	*n* (%) of posts with at least one stigma code (*n* = 268)
Included media (videos, gifs, photos)	138 (19.4)	44 (16.2)
Retweet	448 (63)	167 (62.3)

https://doi.org/10.3768/rtipress.2023.op.0087.2305

RTI International is an independent, nonprofit research institute dedicated to improving the human condition. We combine scientific rigor and technical expertise in social and laboratory sciences, engineering, and international development to deliver solutions to the critical needs of clients worldwide.

www.rti.org/rtipress **RTI Press publication OP-0087-2305**