Contents lists available at ScienceDirect

Child Abuse & Neglect

journal homepage: www.elsevier.com/locate/chiabuneg

Using social media data for assessing children's exposure to violence during the COVID-19 pandemic

Pouria Babvey [a], Fernanda Capela [a], Claudia Cappa [b,*], Carlo Lipizzi [a], Nicole Petrowski [b], Jose Ramirez-Marquez [a]

[a] *Stevens Institute of Technology, 1 Caste Point Terrace, Hoboken, NY 07030, USA*
[b] *UNICEF, Data and Analytics Section, 3 UN Plaza, New York, NY 10017, USA*

ARTICLE INFO

Keywords:
Children
Violence
Abuse
COVID-19
Social media
Big data
Cyberbullying

ABSTRACT

Background: The COVID-19 pandemic brought unforeseen challenges that could forever change the way societies prioritize and deal with public health issues. The approaches to contain the spread of the virus have entailed governments issuing recommendations on social distancing, lockdowns to restrict movements, and suspension of services.
Objective: There are concerns that the COVID-19 crisis and the measures adopted by countries in response to the pandemic may have led to an upsurge in violence against children. Added stressors placed on caregivers, economic uncertainty, job loss or disruption to livelihoods and social isolation may have led to a rise in children's experience of violence in the home. Extended online presence by children may have resulted in increased exposure to abusive content and cyberbullying.
Participants and setting: This study uses testimonial-based and conversational-based data collected from social media users.
Methods: Conversations on Twitter were reviewed to measure increases in abusive or hateful content, and cyberbullying, while testimonials from Reddit forums were examined to monitor changes in references to family violence before and after the start of the stay-at-home restrictions.
Results: Violence-related subreddits were among the topics with the highest growth after the COVID-19 outbreak. The analysis of Twitter data shows a significant increase in abusive content generated during the stay-at-home restrictions.
Conclusions: The collective experience of the COVID-19 pandemic and related containment measures offers insights into the wide-ranging risks that children are exposed to in times of crisis. As societies shift towards a new normal, which places emerging technology, remote working and online learning at its center, and in anticipation of similar future threats, governments and other stakeholders need to put in place measures to protect children from violence.

* Corresponding author.
E-mail addresses: pbabvey@stevens.edu (P. Babvey), fcapela@stevens.edu (F. Capela), ccappa@unicef.org (C. Cappa), clipizzi@stevens.edu (C. Lipizzi), npetrowski@unicef.org (N. Petrowski), jmarquez@stevens.edu (J. Ramirez-Marquez).

https://doi.org/10.1016/j.chiabu.2020.104747
Received 8 August 2020; Received in revised form 11 September 2020; Accepted 13 September 2020
Available online 17 September 2020

1. Introduction

1.1. *Victimization risks faced by children during COVID-19*

The COVID-19 pandemic brought unforeseen challenges that could forever change the way societies prioritize and deal with public health issues. The approaches to contain the spread of the virus across the world have been wide-ranging. Some common approaches have entailed governments issuing recommendations on social distancing for parts or all of their countries as well as suspension of services, from education and childcare to child protection. In other cases, lockdowns were declared with restrictions for all non-essential movements.

Among other concerns, the measures adopted by countries in response to the pandemic may have led to a surge in certain forms of violence against children. The suspension of schools has placed an added burden on families as they struggle to balance childcare, education and work responsibilities. The effects of economic pressure due to reduced or negative economic growth and associated insecurity and uncertainty, job loss, or disruption to livelihoods may increase family violence. Confinement measures have resulted in families spending more time together, which could lead to heightened tensions in the household. All of this, in combination with social isolation, may mean that children face an increased risk of experiencing violence at home (Bradbury-Jones & Isham, 2020; Lee & Ward, 2020; The Alliance for Child Protection in Humanitarian Action, 2020).

Other factors may also have an impact, including that children spend more time online during the pandemic (World Childhood Foundation et al., 2020). Due to stay-at-home orders, schools needed to provide online education and children have been required to use applications and digital platforms. Meanwhile, as they were unable to meet friends in person, children have connected via text messages and social media more often. When children are using educational platforms that demand interactions through posts and comments and they are more connected with peers online, the opportunity for cyberbullying and other forms of online violence increases.

Confinement and disruption of child protective services could cause children who experience violence at home and online to suffer in silence. During the crisis, it is likely more difficult to identify children at risk as many adults who would typically identify signs of abuse and maltreatment (e.g., teachers, childcare workers, coaches, extended family, community members, child and family welfare workers) are no longer in regular contact with children. Therefore, practitioners and institutions that work with children would benefit from analyses of available data that can offer insights into how children's experience of certain forms of violence has changed during the pandemic.

1.2. *Using social media data as a source of information on violence*

Most of the current evidence on the impact of the pandemic on violence against children comes from service providers (Baron, Goldstein, & Wallace, 2020; Benson, Fitzpatrick, & Bondurant, 2020; Peterman, O'Donnell, & Palermo, 2020), and as such cannot be used to assess changes in actual experience of violence. Indeed, data gathered from authorities and those in charge of providing services to children and their families only document cases that are reported to, and reached by, prevention and response mechanisms (United Nations Children's Fund, 2020). These represent only a small portion of all episodes and victims of violence (United Nations Children's Fund, 2014).

Prevalence data, derived from population-based surveys, are needed to obtain representative estimates on the number and characteristics of children who experience violence and to assess trends over time. However, such surveys have been put on hold in many countries due to COVID-19 restriction measures. Additionally, due to safety concerns for victims and researchers as well as methodological constraints, surveys that ask direct questions on the experiences of violence are not advisable during the pandemic (Bhatia, Peterman, & Guedes, 2020; UN Women & World Health Organization, 2020).

Even prior to the pandemic, large-scale prevalence studies of children's exposure to violence have been scant and statistics on this topic have remained inconsistent in scope and quality (Cappa & Petrowski, 2020). While data on violence at home have increased significantly over the last 15 years, forms of abuse that are particularly challenging to measure, such as commercial sexual exploitation, have been largely ignored in data collection (Cappa & Petrowski, 2020). Other topics, such as online abuse, have almost exclusively been investigated in high-income countries (United Nations Children's Fund, 2014). A survey of 44 countries in Western Europe and Canada (conducted between 2017 and 2018) found that the proportion of adolescents who had been cyberbullied varied widely among the participating countries, from 3 percent among 15-year-old boys in Spain to 29 percent among 15-year-old boys in Lithuania (Inchley et al., 2020).

Given the constraints associated with survey data, social media can be leveraged to gather insights into children's well-being and their exposure to violence. Twitter is one such source of information. This microblogging and social networking service allows users to post and interact with messages known as tweets, which can have a maximum length of 280 characters. Unlike most social networks that require users to be at least 13 years of age, children of any age can sign up to Twitter. Additionally, the content posted on Twitter is available and visible to all users, without network restrictions. According to Statista, as of July 2020, there were approximately 330 million Twitter users, of whom 8 percent were below the age of 18. The United States of America (USA), Japan, India, Brazil, the United Kingdom (UK) and Turkey have the greatest number of users on Twitter with 62, 49, 17, 15, 15, and 12 million users, respectively. The official website of Twitter notes that policies are in place to address abusive behaviors and regulate age screening for the purpose of filtering advertisements. Abusive behaviors are defined as any attempt to harass, intimidate, or silence someone else's voice. Abusive or harmful content can be reported and users who are found to engage in such content may have their posts censored or accounts suspended, depending on the gravity of the offense.

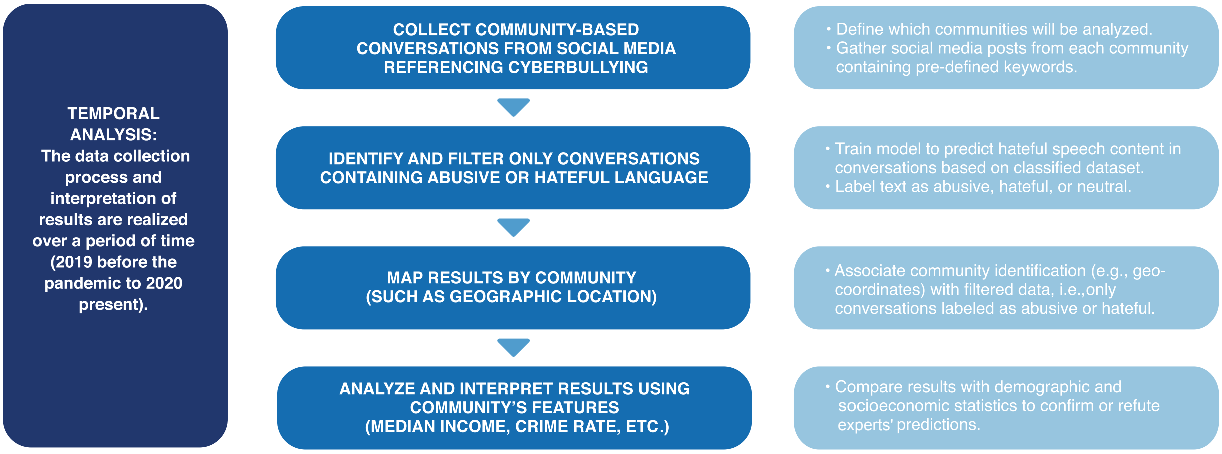
TEMPORAL ANALYSIS: The data collection process and interpretation of results are realized over a period of time (2019 before the pandemic to 2020 present).
COLLECT COMMUNITY-BASED CONVERSATIONS FROM SOCIAL MEDIA REFERENCING CYBERBULLYING
• Define which communities will be analyzed.
• Gather social media posts from each community containing pre-defined keywords.
IDENTIFY AND FILTER ONLY CONVERSATIONS CONTAINING ABUSIVE OR HATEFUL LANGUAGE
• Train model to predict hateful speech content in conversations based on classified dataset.
• Label text as abusive, hateful, or neutral.
MAP RESULTS BY COMMUNITY (SUCH AS GEOGRAPHIC LOCATION)
• Associate community identification (e.g., geo-coordinates) with filtered data, i.e.,only conversations labeled as abusive or hateful.
ANALYZE AND INTERPRET RESULTS USING COMMUNITY'S FEATURES (MEDIAN INCOME, CRIME RATE, ETC.)
• Compare results with demographic and socioeconomic statistics to confirm or refute experts' predictions.

Fig. 1. Approach to assessing change in cyberbullying and abusive or hateful content on Twitter.

Of particular value in the understanding of what topics are most popular on the web are moderated forums that allow discourse between users. Reddit is a social media site with a wide range of forums dedicated to various topics, called subreddits, each of which is moderated by community volunteers. According to Reddit's user agreement, children under the age of 13 are not allowed to create an account or otherwise use the services, even though there is no real way to enforce that requirement. As of 2020, there were over 430 million monthly active Reddit users and Reddit had over 1.2 million different subreddits (Djordjevic, 2020). Among all the users, 7 percent are under age 18 (Response Agency, 2014). For subreddits dedicated to sensitive topics such as depression, domestic abuse, and suicide, the moderators tend to ensure that the anonymous submitter has access to local help hotlines if a life-threatening situation is described. They also enforce respectful behavior by deleting disrespectful or off-topic posts (Schrading, Alm, Ptucha, & Homan, 2015). Users can also report abusive content, which is evaluated by the platform administrators. Unlike Twitter, Reddit allows lengthy submissions, which grant opportunities to cover sensitive subjects that may not typically be discussed in social media. This also makes it possible to apply natural language processing tools for the analysis of the testimonials.

The use of social media platforms to report experiences of violence has grown over the years, alongside academic literature that uses such sources to understand patterns of victimizations. In 2014, Twitter users unequivocally reacted to the assault scandal of Ray Rice, a former American football running back, by unleashing personal stories of domestic abuse via the hashtags #WhyIStayed or #WhyILeft. Schrading, Alm, Ptucha, and Homan (2015) used natural language processing models to extract micro-narratives of both staying and leaving. These included reference to cognitive manipulation, financial constraints, keeping families united, and experiencing shame for #WhyIStayed; and fearing physical violence, realizing self-worth, gaining support, and gaining agency for #WhyILeft. Another major case occurred in 2016, when the Twitter hashtag #MaybeHeDoesntHitYou triggered an outpouring of victims' stories detailing their personal experience of abuse (McCauley, Bonomi, Maas, Bogen, & O'Malley, 2018).

Researchers have also worked on applying tools for automatically detecting reports of violence on social media. In a study by PettyJohn, Muzzey, Maas, and McCauley (2019), a dataset of Facebook posts about domestic violence was collected and manually classified into five categories: awareness, empathy, fund-raising, personal story, and general. According to the results, machine learning models could efficiently label the post into one of the above-mentioned categories. Several studies have also worked on developing models for automatically tagging hate speech and other abusive content on social media. These models mainly used data from Twitter, manually classifying the tweets into abusive and non-abusive using crowdsourcing, and finally training a model for further classifications (Davidson, Warmsley, Macy, & Weber, 2017; Founta et al., 2018; Ousidhoum, Lin, Zhang, Song, & Yeung, 2019; Waseem & Hovy, 2016).

Technology-facilitated abuse has also been recognized as a new form of violence and refers to controlling, monitoring and harassing behaviors using tools such as mobile phones, email, tracking apps and social media (Douglas, Harris, & Dragiewicz, 2019). The frequency and nature of abusive behaviors suggest this is a key form of abuse deserving more significant attention (Douglas et al., 2019). Reddit has been studied less in this area, with work mainly focusing on mental health. In Pavalanathan and De Choudhury (2015), a number of subreddits focusing on mental health topics were identified and used to determine the differences in discourse between throwaway and regular accounts.

1.3. The current study

This study proposes a framework to assess changes in children's exposure to violence during the COVID-19 pandemic using information from social media. It utilizes data collected from Twitter users in 16 countries, as well as information gathered from Reddit in one country. Conversations on Twitter were reviewed to detect cases of abusive or hateful content and cyberbullying, while testimonials from Reddit forums were examined to monitor changes in references to domestic violence and child physical abuse during the pandemic. In addition, the paper discusses the implications of using such information to make inferences about the impact of stay-at-home restrictions on violence against children.

2. Methods and data sources

2.1. Twitter

Fig. 1 shows the process scheme, composed of sequential stages, that was used to assess changes in children's exposure to abusive or hateful content and cyberbullying via conversational social media interactions on Twitter during stay-at-home restrictions. Twitter was chosen for this study as it is one of the most widely used social media platforms in the world and its microblog format permits public conversations between users. Unlike Instagram, Snapchats and Facebook, which rank higher among young users (Chen, 2020), Twitter does not rely as much on videos and images, as engagement mostly happens through textual posts and reposts. Additionally, Twitter has an open data policy that allows for the easy collection of large samples of exchanges filtering by keywords, time stamp, and/or geographic location.

The first unit of the analysis considers users across the 50 states in the USA. The sampling frame includes all the users (1) who have a verified account with a location publicly listed in their profiles; (2) whose location is limited to the USA boundaries; (3) who have posted at least one tweet (or a reply to another tweet) in the time interval between January 2019 to present; (4) and whose tweet or response contains one or some abusive or hateful keywords. The USA was selected for the first unit because at the end of March 2020, the country became the epicenter of the COVID-19 pandemic. By July 2020, the Centers for Disease Control and Prevention was reporting that the country had surpassed the mark of four million coronavirus cases, which was over 25 percent of the total cases worldwide. The first stay-at-home restrictions that ordered the closure of stores and schools came into effect in the first week of March; however, a wide range of measures were implemented across the different states at different times, which did not always mirror the evolution of the pandemic (Lee, 2020).

The first stage of the process entailed the selection of a vast number of textual conversations containing abusive or hateful content, or related to cyberbullying, using a pre-trained model on a labeled dataset to detect the tweets with abusive or hateful content. Abusive language refers to any strongly impolite, rude or hurtful language using profanity, that can show a debasement of someone or something, or conveys intense emotion (Founta et al., 2018; Park & Fung, 2017; Papegnies, Labatut, Dufour, & Linares, 2017). Hateful messages express hatred toward a targeted individual or group, or is intended to be derogatory, to humiliate, or to insult the members of the group, on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender (Davidson et al., 2017). Cyberbullying can be defined as the use of force, threat, or coercion to abuse, embarrass, intimidate, or aggressively dominate others, using electronic forms of contact. It typically denotes repeated and hostile behavior exhibited by a group or an individual (Chatzakou et al., 2017). The conversations were labeled as non-abusive/normal, abusive, hateful, or neutral/uncertain using a Machine Learning algorithm that was trained from a pre-classified dataset. Only tweets containing abusive or hateful language were retained. It is worth mentioning that the classifier found the conversations that contained abusive language among users and did not consider posts in which users discuss their concerns about cyberbullying, potential solutions, etc. Users in social media can be individuals posting about their lives, or institutions, such as newspapers, sharing the latest news and articles, etc. Posts containing the word cyberbullying are likely to be about cyberbullying, but do not necessarily have bullying content. Thus, the classifier labels the news articles, academic and institutional studies as normal. This is an important filter as increases in tweets may be due to more conversations about a topic, rather than an increase in experiences of an issue. For instance, between March 2019 and April 2020, this study collected tweets containing the keyword cyberbullying from New York City. The weekly citations of cyberbullying increased from an average of 20 to 170 citations around March 2020, when the United States started to implement stay-at-home restrictions (*results not shown*). This substantial spike indicates that cyberbullying became a trending topic once the stay-at-home restrictions started. However, this does not mean that there were more cases of cyberbullying in New York City, as these citations can include news articles, experts' predictions, institutional analysis, academic papers, parents' concerns, etc. Indeed, the filtering model found that on average, only 30 percent of the tweets contained hateful or abusive language.

During this process, the metadata containing the sources and dates of the posts were maintained. Once the desired posts were filtered, they were reorganized by date (before and after the stay-at-home restrictions) and by geographic location. The metric used to assess the increase or decrease in abusive tweets was the number of posts per week normalized by the size of the population in the relevant location. Then, these weekly counts were averaged for dates before and after the stay-at-home restrictions started.

To extend the geographic scope of the analysis, 15 more countries around the world were selected as case studies for the second unit of analysis. The sampling frame includes all the users (1) who have a verified account with a location publicly listed in their profiles; (2) whose location is limited to the boundaries of a set of 15 selected countries; (3) and who posted at least one tweet (or a reply to another tweet) in two different timeframes before and after the stay-at-home restrictions (November-December 2019 and March-April 2020). The countries were chosen to provide geographic coverage as well as varying levels of human development as rated by the Human Development Index (HDI) (United Nations Development Programme, 2019). The HDI is an index (between 0 and 1) that measures key dimensions of human development. The three key dimensions are: life expectancy, access to education, and Gross National Income per capita adjusted for the price level of the country. The Coronavirus Government Response Tracker and related Stringency Index (SI), developed by the University of Oxford, were also used to account for the severity of the containment measures in place across countries. This was then compared with the ratio of abusive posts. SI is an index (between 0 and 100) representing 17 indicators of government responses. Eight of the policy indicators (C1-C8) record information on containment and closure policies, such as school closures and restrictions in movement. Four of the indicators (E1-E4) record economic policies, such as income support to citizens or provision of

Table 1
Number of Reddit members and posts by subreddit.

Subreddit	Members in June 2020	Collected posts in 2019/2020
r/abuse	17,000 (rounded)	7,885
r/survivorsofabuse	16,000 (rounded)	4,806
r/domesticviolence	11,000 (rounded)	4,722

Note: The number of members was obtained from www.reddit.com.

Table 2
Main topics in abuse-related subreddits and most frequently used keywords.

Topic	Keywords
Intimate partner abuse	relationship, sex, feel, kiss, abusive, manipulate, friendship, partner, life, date, passive, emotionally, anxiety, advantage, pressure, PTSD, therapy
Physical abuse	grab, car, door, throw, grab, punch, face, neck, bleed, arm, hit, slam, choke, remove, glass, smash, dish, harm, bruise, weed, couch, bed, apartment, floor, wall, hotel
Sexual abuse	sexual, touch, uncomfortable, rape, weird, girl, porn, happen, naked, assault, nude
Child abuse	dad, mom, stepdad, biological dad, stepmom, brother, sister, parent, school, teen, alcoholic, drink, spank, beat
Practitioners' support	hotline, domestic abuse, violence, survivor, legal, community, mail, survey, report, program, protection, court, lawyer, resources, anticipate, payment, anonymous, mistreatment

foreign aid. Five of the indicators (H1-H5) record health system policies such as the COVID-19 testing regime or emergency investments into healthcare. For the purpose of this study, the SI scores of each country were divided by 100 to obtain a value between 0 and 1, and scores were reported as of April 1, 2020. For example, the Philippines, which experienced prolonged and strict lockdowns, had at that point in time an SI of 100, while Sweden, which adopted a much more relaxed approach, had an SI equal to 41.

To measure the possible growth in the number of abusive messages, all the tweets posted from users in a group of 15 different countries during November and December 2019, as well as March and April 2020, were gathered. Over 40 million tweets from users of all ages were collected in total. The collected tweets did not reflect all the tweets posted in the selected countries over the observed time period and only included the tweets posted by verified users and from a location that was publicly listed in the profiles.

A scraping tool was used to gather the tweets, which were later merged into a dataset. A multilingual deep learning model XLM-R, developed in 2019 by Facebook (Lample & Conneau, 2019; Wu, Conneau, Li, Zettlemoyer, & Stoyanov, 2019), was used to aggregate the textual knowledge in different languages. To train the model for abusive message detection, 100,000 tweets published in a study from Founta et al. (2018) were used. Each tweet in that study was classified by a group of five individuals into one of the following categories: normal, abusive, spam, and hateful. A majority vote was then used to decide the final label for each tweet. Fifty-five percent were labeled as normal, 26 percent as abusive, 14 percent as spam, and 5 percent as hateful. After training and tuning, the model was applied to our dataset and tweets were classified into one of the four categories listed above. The model was able to predict the labels of tweets with a 77 percent accuracy and with Cohen's kappa equal to 0.64 (*results not shown*). Only posts classified as hateful or abusive were considered.

To evaluate the performance of the model on languages other than English, a set of 1,000 tweets in Spanish, French, Indonesian, and Arabic (250 tweets from each language) were selected randomly and then manually classified into one of the four mentioned categories. The model was able to label the tweets correctly with 76 percent accuracy compared to 25 percent random assignment (*results not shown*). Although the model showed a reliable performance in detecting the abusive messages, it may have some biases between different languages. However, in this study, the ratio of abusive content between different languages is not used and the analysis relies on comparing the abusive content ratio for each country at different periods of time.

The last stage entailed the compilation of other data to evaluate which features might be crucial to influencing the changes in abusive content. A correlation analysis was performed for the 15 countries between changes in abusive content and human development, as measured by the HDI.

2.2. Reddit

For Reddit, a testimonial-based approach was used to measure changes in violence-related conversations during the pandemic. While Reddit does not rank among the most commonly used social media platforms by children, it was selected in light of the lengthy texts that users can submit, which allows for more refined content analyses. The sampling frame of the analysis on Reddit includes all users who submitted at least one post to one of the subreddits listed in Fig. 7 during the data collection time interval of the 18 months between January 2019 and July 2020. As the majority of the Reddit users, specifically in the violence-related subreddits, are from the USA, the results may not be generalized to the international level.

As a first step, subreddits that focused on family violence were selected. Family violence is defined here as any form of abuse committed by one family member against another, including cases of child maltreatment as well as children's exposure to intimate partner violence. Three abuse-related subreddits were selected, i.e., abuse, survivors of abuse, and domestic violence. Table 1 shows the number of users that were engaged in the three subreddits as of June 2020, as well as the total number of abuse-related posts that were found in 2019 and 2020.

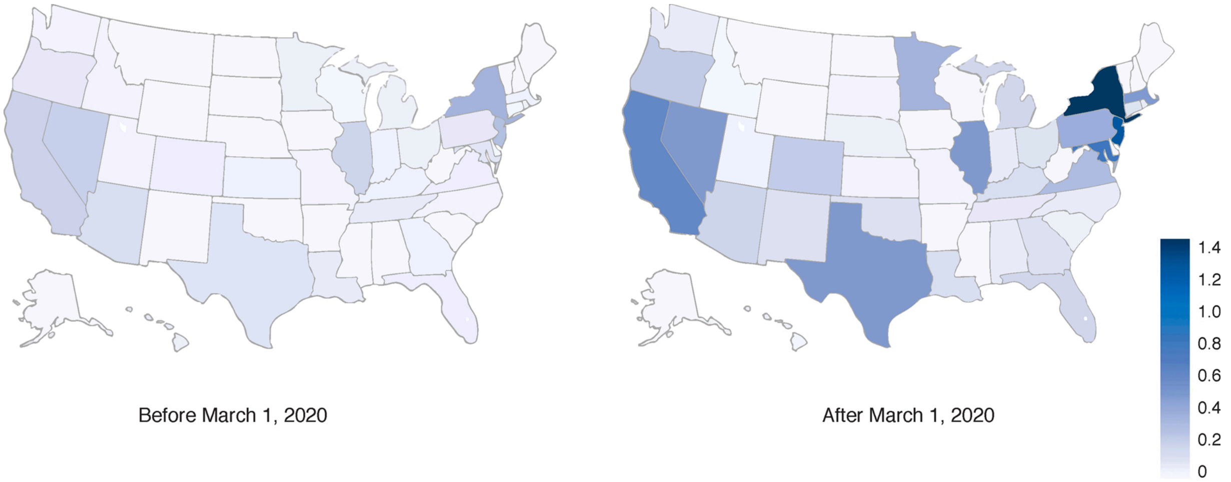

Fig. 2. Weekly number of abusive tweets per 100,000 population before and after the stay-at-home restrictions started by state.
Note: The values for the District of Columbia are not presented in the maps.

Table 3
Average weekly abusive tweets before and after the stay-at-home restrictions started in the top 20 states with the largest number.

	State	Before	After	Increase
1	District of Columbia	0.26	3.55	3.29
2	New York	0.32	1.36	1.04
3	New Jersey	0.26	1.19	0.93
4	Maryland	0.06	0.78	0.71
5	Massachusetts	0.04	0.59	0.55
6	California	0.16	0.60	0.44
7	Nevada	0.16	0.48	0.32
8	Texas	0.08	0.40	0.32
9	Pennsylvania	0.06	0.36	0.30
10	Illinois	0.14	0.45	0.30
11	Minnesota	0.04	0.34	0.30
12	Virginia	0.04	0.30	0.26
13	Colorado	0.04	0.23	0.19
14	Oregon	0.06	0.23	0.17
15	Michigan	0.04	0.17	0.14
16	Florida	0.04	0.18	0.13
17	Oklahoma	0.01	0.12	0.12
18	New Mexico	0.01	0.11	0.11
19	Kentucky	0.03	0.13	0.10
20	Arizona	0.10	0.19	0.09

Overall, 17,413 posts from users of all ages were collected. All the posts were submitted from the beginning of 2019 until the end of June 2020. A topic modeling algorithm was developed to divide the posts in the three violence-related subreddits into five categories (see Table 2). Of all the posts, 42 percent were about intimate partner abuse, 22 percent about physical abuse, 18 percent about sexual abuse, 11.5 percent about child abuse, and 6.5 percent about practitioners' support. One main distinction between physical abuse and intimate partner abuse here is that the model assigned physical abuse to the posts that were reporting physical acts, while it assigned intimate partner abuse to the cases that mainly discussed cognitive manipulation, financial strain, etc.

Additionally, subreddits dedicated to sensitive topics such as depression, suicide, relationship disorders, etc. were added as control groups. Comparing violence reports with other types of reports allows for a more accurate estimation of changes, as more people may have used social media as a medium to share their personal experience during the stay-at-home restrictions when other forms of interaction were limited.

Python Reddit Application Programming Interface (API) Wrapper (PRAW) was used to extract the data. During the pre-processing step, links, usernames, numbers, and stop words (i.e., most common, short function words, such as: the, is, at, out) were removed. Also, all the posts were normalized by lowercasing and lemmatizing (Balakrishnan & Lloyd-Yemoh, 2014). Topic modeling was used to extract the posts, categorize the posts into different types of abuse, and measure the share of each topic before and after the stay-at-home restrictions. Topic modeling algorithms categorize a set of posts into a number of different groupings. In this study, the Latent Dirichlet Allocation (LDA) algorithm was used for topic modeling. LDA (Blei, Ng, & Jordan, 2003) was developed based on

Table 4
Ngrams extracted from abusive tweets before and after the stay-at-home restrictions started.

Ngrams before March 1, 2020		Ngrams after March 1, 2020	
fucking_wheezing	badly_hate	feel_bad	mad_michelle
criminal_behavior	bullies_learn	wrong_calling	got_wedgie
real_corny	just_gross	**cyberbullying_animal**	rotten_bitch
literally_grow	lies_everyday	**stop_whining**	shaming_dumbasses
stupid_drunk	cyberbullying_fun	phone_cyberbullying	know_shit
ima_report	like_wtf	**cyberbullying_myth**	dude_stop
facebook_blocked	aaaaaaannnnnnnnd_ban	**cyberbullying_sanctified**	**cyberbullying_works**
kelly_afraid	dumb_fuck	**faccio_cyberbullying**	**cyberbullying_cyberhyping**
fucking_idiot	stop_cyber_bullying	ugh_gross	karma_hoe
demented_trick	fuck_cyber_bullying	cyberbullying_asap	really_disgusts
scrappy_doo	burn_you_alive	**just_shitposting**	stop_shaming
try_better	bullying_that_poor	attempt_suicide	cyberbullying_bad
cyber_yikes	retweeting_paticularly_nasty	fucking_ruthless	quarantine_cyber_bullying

Note: The expressions in bold denote the tweets that defend cyberbullying or even show pride in the action.

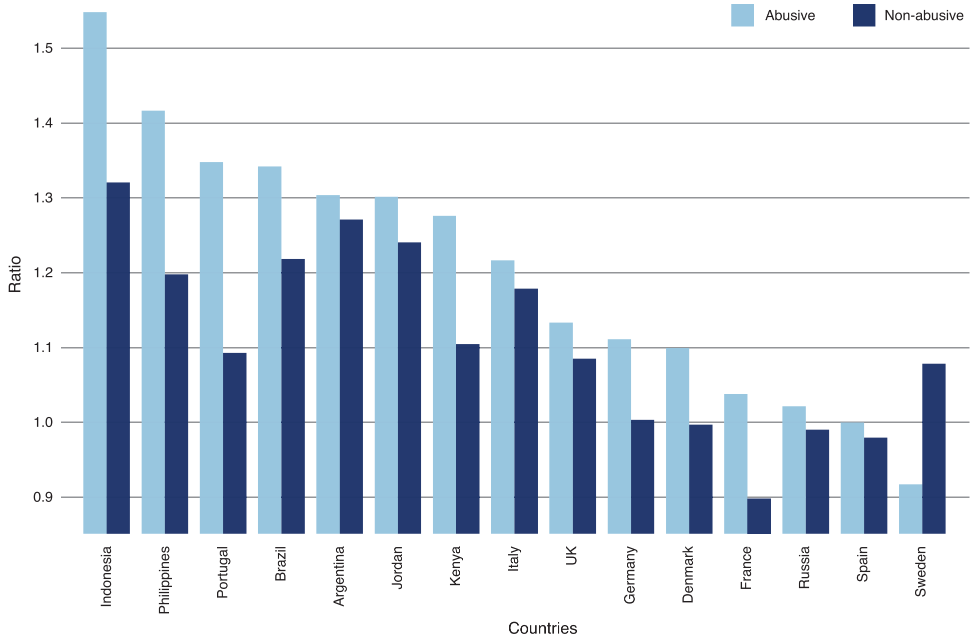

Fig. 3. Ratio between the number of tweets from November-December 2019 and March-April 2020 by whether the tweets were abusive or non-abusive.
Note: A ratio above 1 indicates an increase in number of tweets.

generative statistical models, and most of the existing topic modeling methods are an extension or a variation of LDA (Chen, Kou, Shang, & Chen, 2015; Wang, Blei, & Heckerman, 2012). LDA draws a distribution over words per topic and a distribution over topics per document. Then, the most frequent words in each topic are selected as keywords. Finally, based on the set of the generated keywords for different categories, a proper topic can be assigned to each category. To calculate the growth of the number of abuse-related reports on Reddit, the daily average number of posts in the selected subreddits before and after the stay-at-home restrictions were compared. March 17, 2020 was used as the beginning of the stay-at-home restrictions in the USA (based on the University of Oxford's Coronavirus Government Response Tracker).

3. Results

3.1. Analysis of Twitter

The analysis of Twitter data shows a significant increase in abusive content generated during the stay-at-home restrictions.

Fig. 2 illustrates abusive tweets collected from the 50 states in the USA before and after March 2020. A spike in such tweets can be observed after March 2020, across different states. As shown in Table 3, the District of Columbia shows the highest increase in abusive tweets, followed by New York.

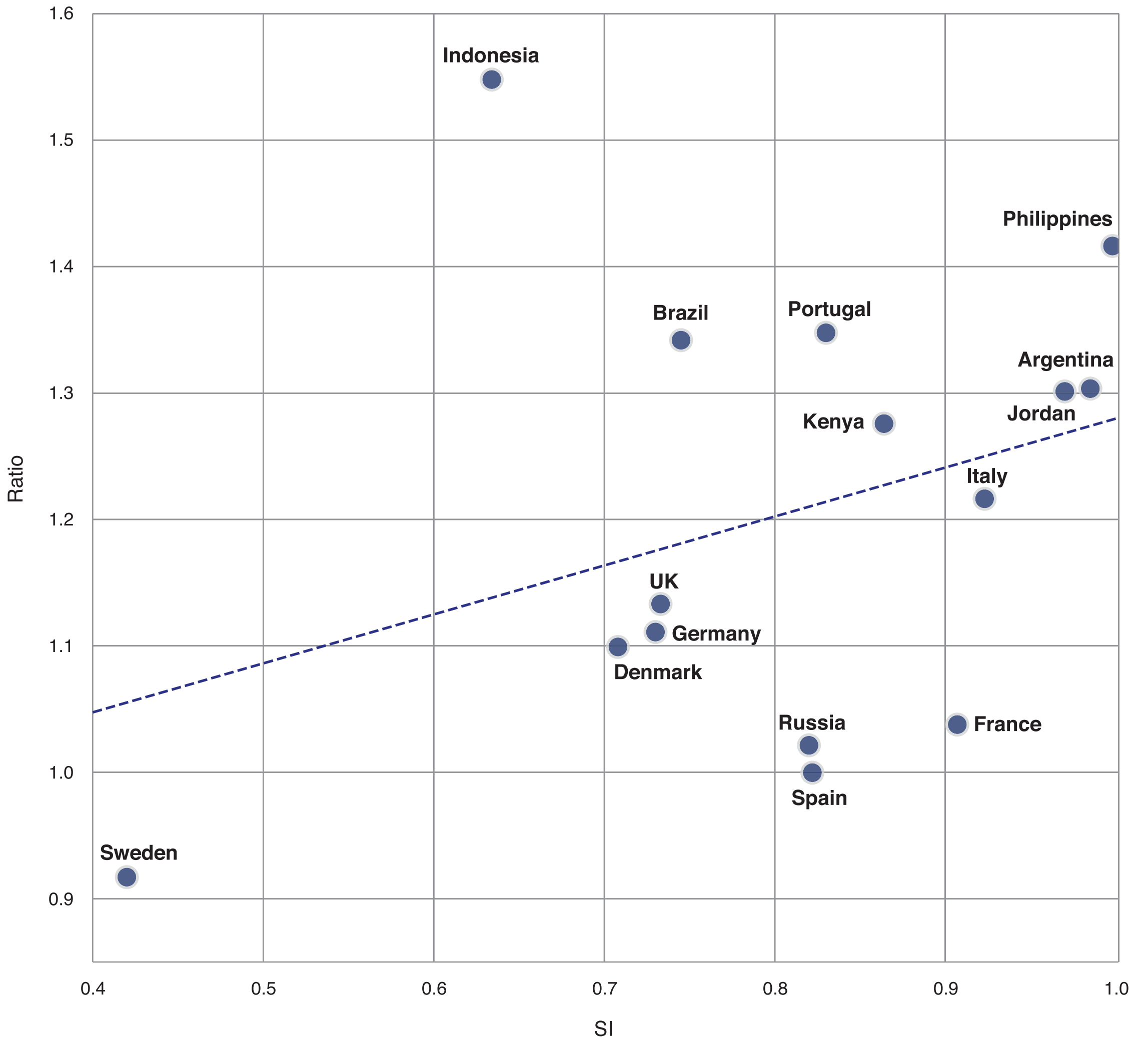

Fig. 4. Ratio between the number of tweets from November-December 2019 and March-April 2020, and the SI.

With the increase of social media usage during the stay-at-home restrictions, it was expected that the number of tweets would increase. However, the overall increase in Twitter usage after the stay-at-home restrictions started has been 24 percent (The Washington Post, 2020), which is much smaller than the average weekly increase in abusive tweets.

Table 4 shows commonly used *ngram* (i.e., a sequence of *n* items from a given text) that were extracted from the abusive or hateful tweets posted in the reference period. It is worth noting that some of the expressions that appeared after the stay-at-home restrictions started, such as those shown in bold in the table, seem to defend cyberbullying or even show pride in the action.

Fig. 3 presents the change in abusive and non-abusive tweets between November-December 2019 and March-April 2020 for the additional 15 countries. The number of abusive messages increased in all the countries, but Sweden. The growth was particularly significant in Indonesia, the Philippines, Portugal, and Brazil, with a more than 30 percent increase compared to 2019. One possible explanation could be that due to the stay-at-home measures more people, especially adolescents, were actively using Twitter. However, Fig. 3 also shows that the growth of abusive content has been more significant than the increase in non-abusive content.

Finally, Figs. 4 and 5 show the growth rate of abusive tweets and the Stringency Index (SI), and the growth rate of abusive tweets and the Human Development Index (HDI), respectively, for the 15 countries. On the one hand, the growth of abusive content was not clearly or strongly correlated with the SI. On the other hand, a negative correlation was found between the growth in abusive content and the HDI scores. Based on the statistical analysis results, the correlation coefficient between the HDI and abusive content growth is 0.7, while the correlation between the SI and abusive content growth is 0.33.

3.2. *Analysis of Reddit*

The analysis of Reddit shows that violence-related subreddits were among the topics with the highest growth after the COVID-19 outbreak.

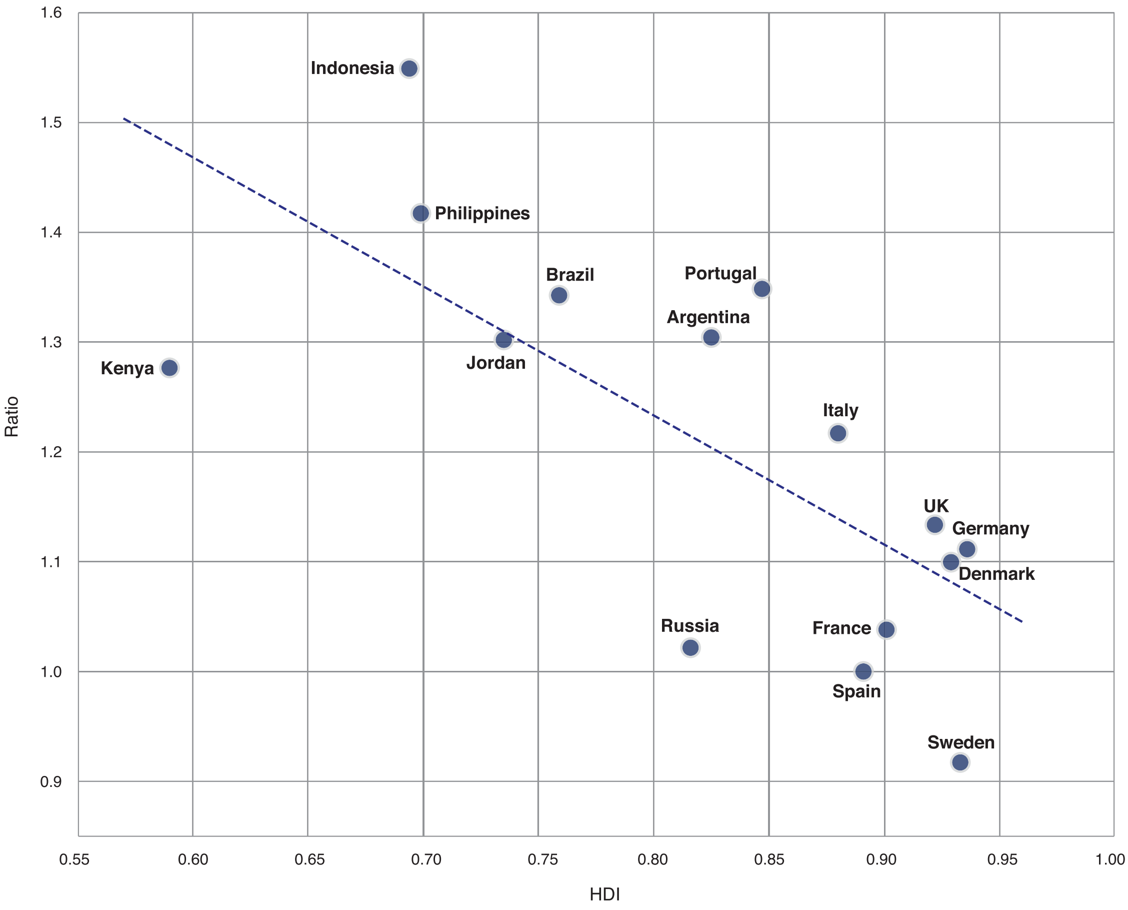

Fig. 5. Ratio between the number of tweets from November-December 2019 and March-April 2020, and the HDI.

Fig. 6 shows the average weekly posts in subreddits on sensitive topics, expressed as a ratio. The three abuse-related subreddits were among those with the largest increase in level of activities after the stay-at-home restrictions started. Topics like r/selfhelp, r/selfharm, r/insomnia, and r/anger also show significant growth, while others like r/socialanxiety, r/stopdrinking, and r/bipolar do not show a significant change.

Fig. 7 shows the ratio in the average daily number of posts before and after the stay-at-home restrictions. After the stay-at-home restrictions started, this figure is almost 35 percent higher than the daily average before the restrictions went into effect. However, for the r/abuse, r/selfhelp and r/survivorsofabuse the growth is more significant, while for others like r/socialanxiety and r/stopdrinking the results actually show a decrease.

To see how much of the increased content in abuse-related subreddits was produced by newly active users, the date of the first message of each user was recorded from the beginning of 2019 until the end of June 2020. Fig. 8 shows the number of newly active users each day in the 18 months beginning from January 2019. The results show a significant increase in the number of users who joined violence-related subreddits after the stay-at-home restrictions started.

Finally, Fig. 9 shows the average monthly number of posts for each of the five categories of abuse described earlier. When comparing the average monthly number of posts between April 2019 and February 2020 to the same number for the period between March 2020 and July 2020, the results show increases in a number of areas. There was a 106 percent growth in posts related to physical abuse, a 94 percent increase for child abuse, an 88 percent increase for intimate partner abuse, and a 62 percent increase for sexual abuse.

4. Discussion

4.1. Implications

The analysis of Twitter and Reddit data shows an increase in abusive or hateful conversations and violence-related testimonials on these social media platforms after confinement measures were enacted. More specifically, abuse-related subreddits were among those with the highest activity growth after March 17, 2020. This was the date that stay-at-home restrictions began in many states within the

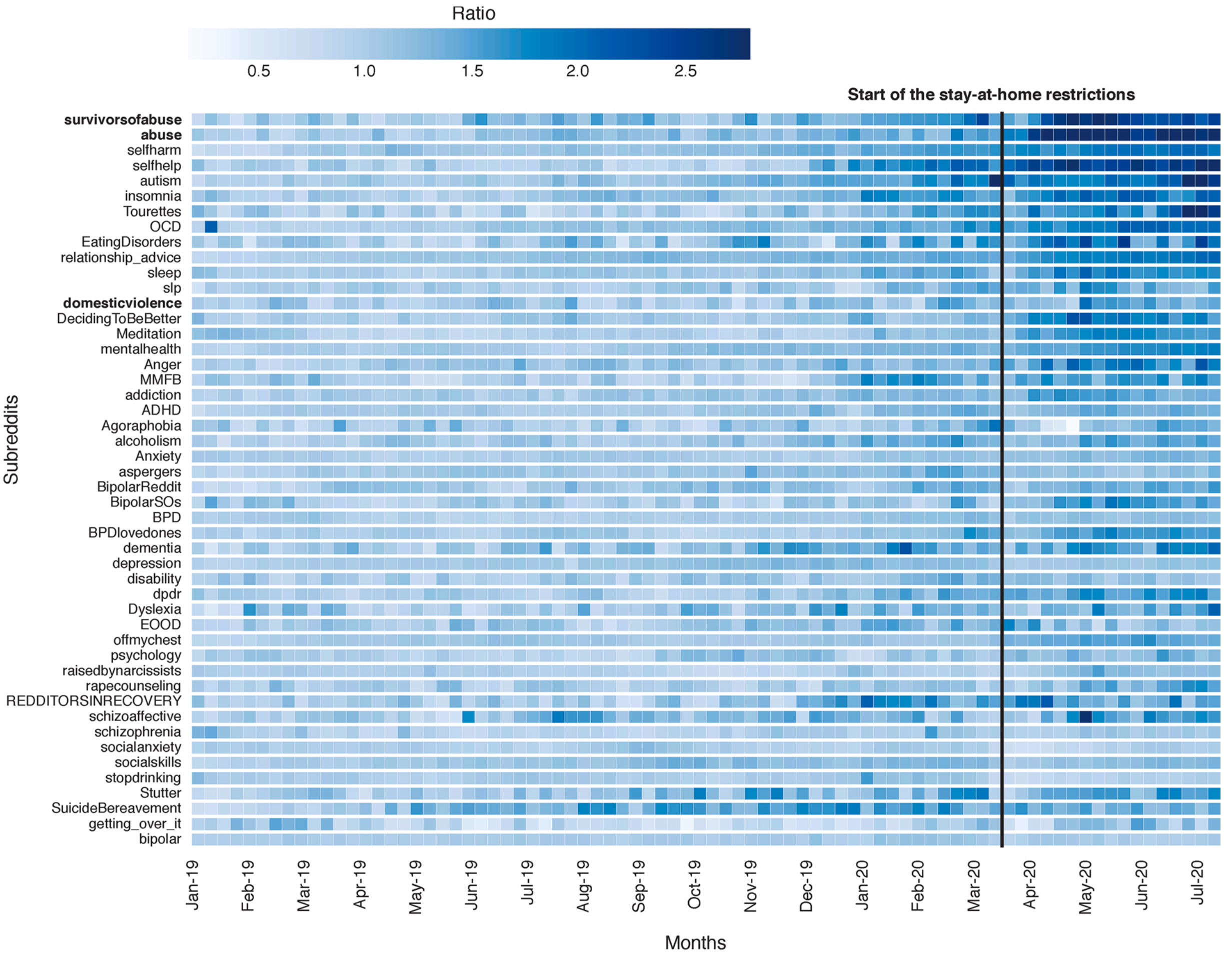

Fig. 6. Ratio of the number of weekly posts to the average in the 18 months between January 2019 and July 2020 for selected subreddits. Note: Abuse-related subreddits are in bold.

USA. The growth in testimonials related to abuse was higher than the growth in posts about other sensitive topics, including mental health. These findings should be considered in light of other research and studies which have found that COVID-19 restriction measures have inhibited the reporting of child maltreatment to authorities and professionals, such as teachers, who are typically in regular contact with children (see for example: Baron et al., 2020). This might suggest that, given the increased online presence of both children and adults, service providers should explore innovative ways through which social media and virtual platforms might be used to identify potential victims of violence, and to consider how these platforms could act as entry points for setting up reporting and referral mechanisms as well as the provision of support services. This, of course, needs to be done in ways that guarantee confidentiality and privacy of the victims and does not expose them to further harm.

Similarly, there was a notable increase in the use of abusive or hateful language on Twitter in the USA after March 1, 2020 as compared to before this period. The observed rise in abusive online content was not restricted just to the USA. The fact that the growth of abusive content during the pandemic was not strongly associated with the Stringency Index implies that increased exposure to harmful content online is occurring across the world regardless of how strict a government's response to COVID-19 has been. This might point to reasons other than the severity of the measures driving the increase in abusive content, including increased concerns and stress deriving from the pandemic and related containment measures and socioeconomic impacts. Regardless of whether or not children themselves are being targeted or victimized, the public nature of the content means that children are at risk of being exposed to such abusive language and negative interactions between online users and this poses a risk to their well-being. This means that response and increased safety mechanisms need to be developed and put into place to prevent children's further exposure to abusive content as the global culture shifts toward an increasingly digital world. ICT companies and social media platforms have an important role to play in keeping children safe online by establishing or strengthening policies and mechanisms for privacy and protection of users from abusive content.

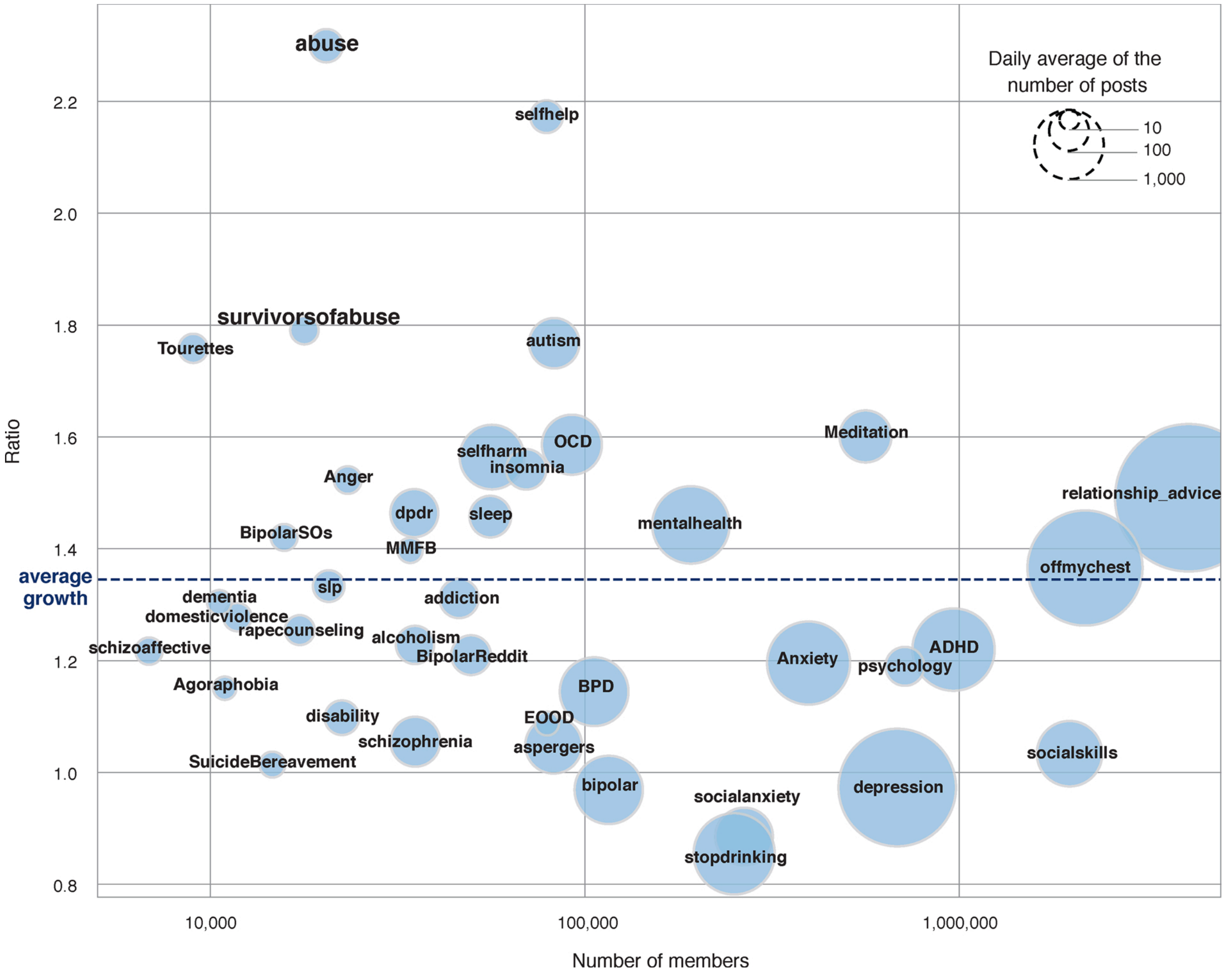

Fig. 7. Ratio of the number of daily posts before and after the stay-at-home restrictions started for selected subreddits.
Notes: The size of the circles is proportional to the average daily number of posts for each subreddit. Small-size and mid-size subreddits are displayed on the left side of the figure and large subreddits are on the right side. OCD stands for obsessive-compulsive disorder; BPD stands for borderline personality disorder; DPDR stands for depersonalization and derealization; EOOD stands for exercise out of depression; SLP stands for speech-language therapy; and MMFB stands for make me feel better.

4.2. *Limitations of the study and directions for future research*

A limitation of the study is that it was not possible to restrict the analyses to only those involving children for two main reasons. The first has to do with privacy and the second is due to limitations with users disclosing their age (or being truthful about their age) in their online profiles. For these reasons, the findings cannot be taken as an indication of children directly experiencing increased cyber-bullying but should rather be interpreted as evidence of the potential for children to be exposed to increasingly abusive content while online. Similarly, the growth in violence-related testimonials on Reddit cannot be interpreted as conclusive of increases in family violence, as this may be due to the extended online presence and lack of other venues where users can talk about their experiences.

Caution is warranted when interpreting the findings from both sets of analyses and the results should not be generalized to other country settings since the majority of Reddit users, and in particular those accessing subreddits about violence, are known to be based in the USA. At the same time, the Twitter analysis was carried out on only 15 countries (in addition to the USA).

The current study demonstrates the potential for using social media data to shed light on children's exposure to violent content online. Future research and additional analysis are needed to zoom in on the specific experiences of children online during the COVID-19 pandemic.

Finally, the approach used in this paper to analyze abuse-related content on social media proved effective in identifying patterns and assessing trends. The application of such an approach to other subjects should be explored.

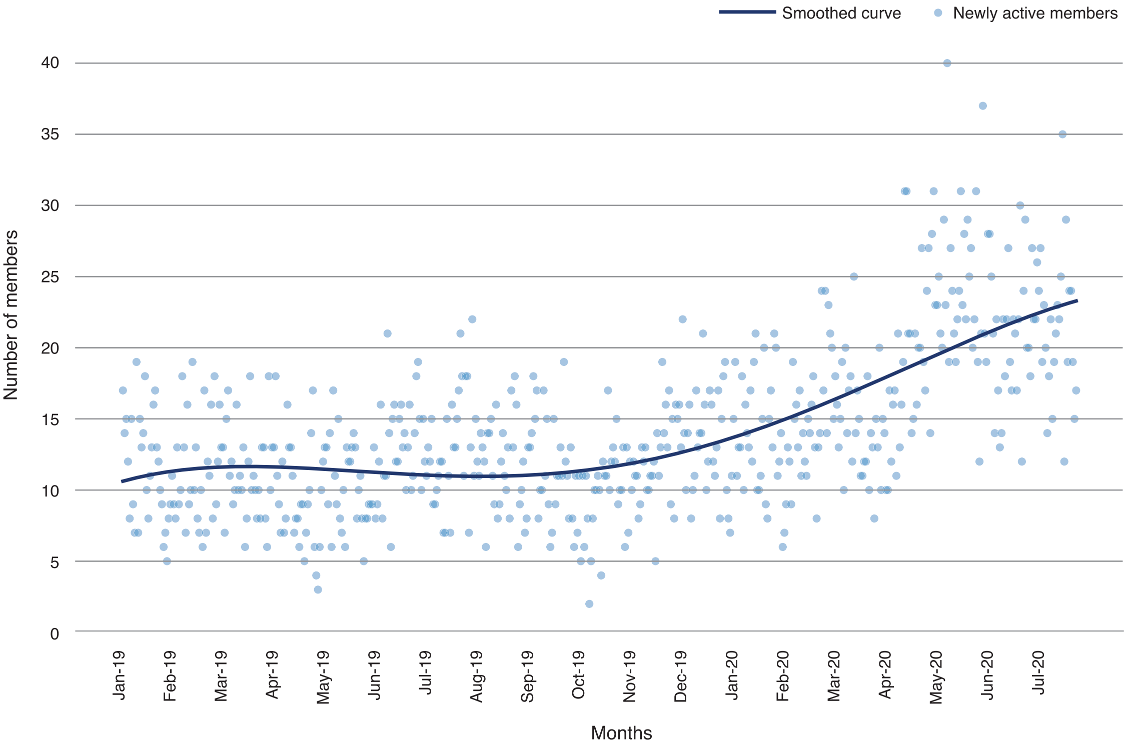

Fig. 8. Number of newly active users on abuse-related subreddits before and after the stay-at-home restrictions started.

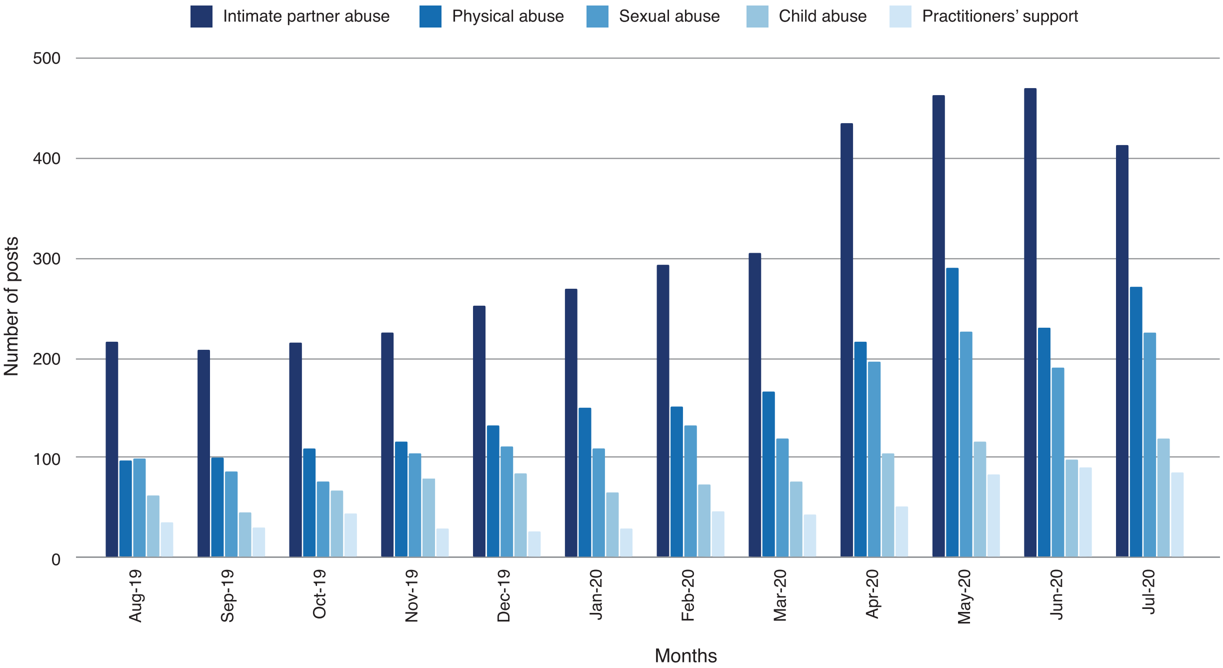

Fig. 9. Average monthly number of posts for each of the five categories of abuse.

5. Conclusions

The COVID-19 pandemic and related containment measures offer insights into the wide-ranging risks that children are exposed to in times of crisis. In documenting an increase in abusive posts on social media during the stay-at-home restrictions, this paper identifies the need for safe and effective remote-access support mechanisms. The growth in violence-related testimonials online sheds light on the role of social media platforms as venues for disclosing experiences of abuse and, possibly, counseling in times when in-person interactions are limited. As societies shift toward a new normal, with technology, remote working and learning at its center (and anticipating similar future threats), governments and other stakeholders need to adopt measures to support families, regulate online platforms and strengthen services as they work to protect children from all forms of violence.

References

Balakrishnan, V., & Lloyd-Yemoh, E. (2014). Stemming and lemmatization: A comparison of retrieval performances. *Lecture Notes on Software Engineering, 2*(3), 262–267. https://doi.org/10.7763/LNSE.2014.V2.134

Baron, J. E., Goldstein, E. G., & Wallace, C. T. (2020). Suffering in silence: How COVID-19 school closures inhibit the reporting of child maltreatment. *Journal of Public Economics, 190*, 04258.

Benson, C., Fitzpatrick, M.D., Bondurant, S. (2020). *Beyond reading, writing, and arithmetic: The role of teachers and schools in reporting child maltreatment. NBER working paper No. w27033*. April.

Bhatia, A., Peterman, A., & Guedes, A. (2020). *Remote data collection on violence against children during COVID-19: A conversation with experts on research priorities, measurement and ethics (Part 2)*. *Innocenti Think Piece*. https://www.unicef-irc.org/article/2004-collecting-remote-data-on-violence-against-children-during-covid-19-a-conversation.html.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. January *Journal of Machine Learning Research, 3*, 993–1022.

Bradbury-Jones, C., & Isham, L. (2020). The pandemic paradox: The consequences of COVID-19 on domestic violence. *Journal of Clinical Nursing, 29*, 2047–2049. https://doi.org/10.1111/jocn.15296

Cappa, C., & Petrowski, N. (2020). Thirty years after the adoption of the Convention on the Rights of the Child – Progress and challenges in building statistical evidence on violence against children. *Child Abuse and Neglect*, 104460. https://doi.org/10.1016/j.chiabu.2020.104460

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Mean birds: Detecting aggression and bullying on twitter. *9th ACM WebScience*.

Chen, J. (2020). *Social media demographics to inform your brand's strategy in 2020*. May 5. SproutSocial https://sproutsocial.com/insights/new-social-media-demographics/#SC-demos.

Chen, K., Kou, G., Shang, J., & Chen, Y. (2015). Visualizing market structure through online product reviews: Integrate topic modeling, TOPSIS, and multi-dimensional scaling approaches. *Electronic Commerce Research and Applications, 14*(1), 58–74.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. May 15-18 *Proceedings of the eleventh international AAAI conference on web and social media*, 512–515.

Djordjevic, N. (2020). *109 Ridiculous Reddit statistics & facts to know in 2020*. January 8. Website Builder https://websitebuilder.org/reddit-statistics/.

Douglas, H., Harris, B. A., & Dragiewicz, M. (2019). Technology-facilitated domestic and family violence: Women's experiences. *The British Journal of Criminology, 59* (3), 551–570.

Founta, A. M., et al. (2018). *Large scale crowdsourcing and characterization of twitter abusive behavior*. arXiv:1802.00393.

Inchley, J., Currie, D., Budisavljevic, S., Torsheim, T., Jåstad, A., Cosma, A., et al. (Eds.). (2020). *Spotlight on adolescent health and well-being. Findings from the 2017/2018 Health Behaviour in School-aged Children (HBSC) survey in Europe and Canada. International report. Volume 1. Key findings*. Copenhagen: WHO Regional Office for Europe.

Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining [Conference paper]. In *33rd Conference on neural information processing systems (NeurIPS 2019)*.

Lee, A. (2020). *These states have implemented stay-at-home orders. Here's what that means for you*. April 7. CNN https://www.cnn.com/2020/03/23/us/coronavirus-which-states-stay-at-home-order-trnd/index.html.

Lee, S. J., & Ward, K. P. (2020). *Stress and parenting during the coronavirus pandemic [Research brief]*. March 26. Parenting in Context Research Lab.

McCauley, H. L., Bonomi, A. E., Maas, M. K., Bogen, K. W., & O'Malley, T. L. (2018). #maybehedoesnthityou: Social media underscore the realities of intimate partner violence. *Journal of Women's Health, 27*(7), 885–891.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., & Yeung, D. Y. (2019). *Multilingual and multi-aspect hate speech analysis*. arXiv preprint arXiv:1908.11049.

Papegnies, E., Labatut, V., Dufour, R., & Linares, G. (2017). Detection of abusive messages in an on-line community. *Conference en Recherche d'Information et Applications*.

Park, J. H., & Fung, P. (2017). *One-step and two-step classification for abusive language detection on twitter*. arXiv preprint arXiv:1706.01206.

Pavalanathan, U., & De Choudhury, M. (2015). Identity management and mental health discourse in social media. May *Proceedings of the 24th International Conference on World Wide Web*, 315–321.

Peterman, A., O'Donnell, M., & Palermo, T. (2020). *COVID-19 and violence against women and children: What have we learned so far?*. June. CGD note. Center for Global Development.

PettyJohn, M. E., Muzzey, F. K., Maas, M. K., & McCauley, H. L. (2019). #Howiwillchange: Engaging men and boys in the #metoo movement. *Psychology of Men & Masculinities, 20*(4), 612–622.

Response Agency. (2014). *Reddit demographics*. February 10. Response Agency https://response.agency/blog/2014/02/reddit-demographics-and-user-surveys/.

Schrading, N., Alm, C. O., Ptucha, R., & Homan, C. (2015a). An analysis of domestic abuse discourse on reddit. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 2577–2583).

Schrading, N., Alm, C. O., Ptucha, R., & Homan, C. (2015b). #whyistayed,#whyileft: Microblogging to make sense of domestic abuse. In *Human language technologies: The 2015 annual conference of the North American chapter of the ACL* (pp. 1281–1286).

The Alliance for Child Protection in Humanitarian Action. (2020). *Technical note: Protection of children during the coronavirus pandemic*. March. version 1. https://alliancecpha.org/en/COVD19.

The Washington Post. (2020). *Twitter usage up, but advertising sales slow*. April 30. The Washington Post https://www.washingtonpost.com/business/economy/twitter-sees-record-number-of-users-during-pandemic-but-advertising-sales-slow/2020/04/30/747ef0fe-8ad8-11ea-9dfd-990f9dcc71fc_story.html.

UN Women and World Health Organization. (2020). *Violence against women and girls data collection during COVID-19. [Brochure]*. April 17 https://www.unwomen.org/-/media/headquarters/attachments/sections/library/publications/2020/vawg-data-collection-during-covid-19-compressed.pdf?la=en&vs=2339.

United Nations Children's Fund. (2014). *Hidden in plain sight: A statistical analysis of violence against children*. New York: UNICEF. https://data.unicef.org/resources/hidden-in-plain-sight-a-statistical-analysis-of-violence-against-children/.

United Nations Children's Fund. (2020). *Strengthening the availability and quality of administrative data on violence against children: Challenges and promising practices from a review of country experiences*. New York: UNICEF.

United Nations Development Programme. (2019). *Human development report 2019*. New York: UNDP.

Wang, C., Blei, D., & Heckerman, D. (2012). *Continuous time dynamic topic models*. arXiv:1206.3298.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. June12-17, *Proceedings of the NAACL-HLT 2016*, 88–93.

World Childhood Foundation, End Violence against Children, ITU, UNESCO, UNICEF, UNODC, et al. (2020). *COVID-19 and its implications for protecting children online*. April.

Wu, A., Conneau, A., Li, H., Zettlemoyer, L., & Stoyanov, V. (2019). *Emerging cross-lingual structure in pretrained language models*. arXiv:1911.01464.

Web references

Statista: https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/, Accessed on August 7, 2020.
Twitter help: https://help.twitter.com/en/rules-and-policies/abusive-behavior, Accessed on August 7, 2020.
Reddit user agreement: www.redditinc.com/policies/user-agreement#section_children_and_reddit, Accessed on August 7, 2020.
Centers for Disease Control and Prevention: www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html, Accessed on August 7, 2020.
University of Oxford's Coronavirus Government Response Tracker: https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker, Accessed on August 7, 2020.