American Behavioral Scientist
2021, Vol. 65(5) 689–711

DOI: 10.1177/0002764221989777
journals.sagepub.com/home/abs


Mapping out Violence Against Women of Influence on Twitter Using the Cyber–Lifestyle Routine Activity Theory

Priya Kumar[1], Anatoliy Gruzd[1], and Philip Mai[1]

Abstract

The study applies and expands the routine activity theory to examine the dynamics of online harassment and violence against women on Twitter in India. We collected 931,363 public tweets (original posts and replies) over a period of 1 month that mentioned at least one of 101 influential women in India. By undertaking both manual and automated text analysis of "hateful" tweets, we identified three broad types of violence experienced by women of influence on Twitter: dismissive insults, ethnoreligious slurs, and gendered sexual harassment. The analysis also revealed different types of individually motivated offenders: "news junkies," "Bollywood fanatics," and "lone-wolves", who do not characteristically engage in direct targeted attacks against a single person. Finally, we question the effectiveness of Twitter's form of "guardianship" against online violence against women, as we found that a year after our initial data collection in 2017, only 22% of hostile posts with explicit forms of harassment have been deleted. We conclude that in the social media age, online and offline public spheres overlap and intertwine, requiring improved regulatory approaches, policies, and moderation tools of "capable" guardianship that empower women to actively participate in public life.

Keywords

gender, online harassment, violence against women, social media, content analysis, social network analysis, India, Twitter

[1]Ryerson University, Toronto, Ontario, Canada

Corresponding Author:
Anatoliy Gruzd, Ted Rogers School of Information Technology Management, Ryerson University, 350 Victoria Street, Toronto, Ontario M5B 2K3, Canada.
Email: gruzd@ryerson.ca

Introduction

In May 2017, prominent Indian journalist Barkha Dutt took to the *Hindustan Times* to share some of the vitriolic, sexist, and abusive messages that she receives daily via Twitter (Dutt, 2017). Dutt's message was simple: for many women, especially those in public roles, receiving online abuses, threats, and expressions of violence are becoming an unfortunate "cost of doing business" in the social media age. With the rapid growth and spread of social media, the number of people exposed to and engaged in acts of online violence, abusive, and antisocial discursive behaviors globally has risen exponentially (Kwak et al., 2015). In the past few years alone, there has been a growing awareness of "toxic" online cultures (such as #Gamergate and The Fappening[1]) with online communities on social media emerging as public hotspots for online misogyny and expressions of violence toward women specifically (Burgess & Matamoros-Fernández, 2016). According to a 2015 report released by the United Nations Broadband Commission's Working Group on Gender, approximately 73% of women across the globe have been targeted or exposed to some form of violence online (e.g., threats, harassment, or stalking). A 2017 Amnesty International IPSOS MORI commissioned poll found that women are more likely to confront and experience mental and psychological trauma due to online harassment, and restrict their content posting thereafter (Amnesty International, 2017). And while there is a growing concern over the social, political, economic, and health-related consequences of online violence against women (VAW), current preventative frameworks and reporting tools offer limited defensive strategies of protective guardianship and long-term solutions to this ever-growing problem. Furthermore, much of the research in this area has been in a Western anglophone context (North America and Europe), omitting women's online experiences from the Global South (Ging & Siapera, 2018).

Through the lens of Habermas (1962/1989, many scholars have theorized social media like Twitter to mimic local community centers, churches, parks, and neighborhoods of the past. That is, social media platforms provide outlets for friendship and communication networks to flourish (Fuchs, 2014; Gruzd & Haythornthwaite, 2013). At the same time, these online public spaces are far from neutral or egalitarian with respect to ensuring the safety, equal representation, and participation of historically marginalized groups and members of society (Correa & Pavez, 2016). Scholars like Shaw (2014) have pointed out that discriminatory biases in tech culture and inequalities observed in today's "digital infrastructures" are also symptoms of a much wider patriarchal society. The firing of former Google engineer James Damore, author of the controversial 10-page "antidiversity" memo is a prime example of the unseen misogyny and sexism that seeps into many professional work environments (Chachra, 2017; Marwick, 2013). Hence, we can see how digital infrastructures including social media platforms often reinforce offline power relations with the concerns of marginalized groups like women being overlooked and disproportionately targeted online by harassers, aggressors, and trolls (Gray, 2012).

The aim of the present study is to better understand the online dynamics of VAW in India through four core objectives. First, the research aims to contribute a greater understanding of online VAW in a non-Western cultural context. Second, the research seeks to gain a better sense of the victims and offenders of online VAW. Third, the research strives to develop the lexicon of online VAW. Finally, the current study aims to contribute new theoretical insights to apply and expand on notions of 'capable guardianship' in the routine activity theory (RAT) to study and develop solutions around the problem of VAW on social media. We conclude by questioning the capabilities and overall effectiveness of Twitter's platform-based guardianship (i.e., automated and human-led content moderation) and discuss current policy responses and knowledge gaps to combatting online VAW.

We specifically study online violence against women of influence. Participating on social media like Twitter has become a norm in modern outreach strategies, and this is especially the case for leaders seeking to maintain public visibility or gain influence. For instance, corporate leaders are increasingly encouraged to actively engage in interpersonal communication strategies like maintaining a social media presence (Neal, 2017). Unfortunately, experiences of online VAW now come with the job for many women with public-facing careers or those in leadership positions in male-dominated fields (Chess & Shaw, 2015; Cooney, 2018). The next section will discuss how we developed our research questions based on the previous work in this area.

Previous Work and Research Questions

Case Study of Twitter Use in India

In this research, we investigate online VAW in India, a country with a population of 1.2 billion where roughly 41% of women have reported to have experienced some form of harassment online (Bhargava, 2017). While recent technological innovations and digital literacy policies (Darade, 2017) have helped propel India forward into the social media age, institutional challenges continue to persist, with issues of gender equality remaining one of the largest. Indian women's experiences of victimization also manifest online and continue to go largely unreported due to the embedded patriarchal structures, fears of family rejection, and societal repercussions that perpetuate gender inequality offline (Amnesty International India, 2018). We argue that without improved systems of prevention and protection, online experiences of gender-based violence will continue to challenge women's safety.

For the purposes of our research, we examine Twitter use in India because of the platform's growing importance for information sharing and news consumption in the country (Malhotra & Malhotra, 2016). Recent reports highlight India as Twitter's fastest growing market, especially from a user-base audience perspective (Mitter, 2015). Twitter's 2017 earnings show that Twitter's daily active users in the Indian market grew at a rate 5x higher than the global average (Chaturvedi, 2017). Twitter has widened its audience since its launch in 2006, and as of January 20, 2021, it ranks 26th in terms of internet traffic and engagement in India (Alexa.com, 2021).[2]

Thus, our first research question is as follows:

Research Question 1: How is online VAW manifested on Twitter in the Indian context?

Victims, Offenders, and Guardians

Next, to guide our research, we turn to the "cyber–lifestyle" RAT. Traditionally, the routine activity theory (RAT) has been used to explain how face-to-face physical interactions can motivate individuals to engage in deviant behavior against victims lacking protective guardianship (Cohen & Felson, 1979). Felson and Eckert (2015) commonly refer to RAT as "the chemistry of crime," positing that transgressions between humans are more likely to occur when the following factors converge in time and space: (a) an "attractive" and accessible target victim, (b) a "motivated" offender, and (c) an absence of a guardian "capable" of intervention (Holt & Bossler, 2008). When these three core components intersect, there is a greater likelihood that an offense such as harmful speech or expressions of violence will take place.

The RAT has been successfully employed to study adolescent cyberbullying and harassment (Reyns et al., 2011), online hate speech (Costello et al., 2016), malware victimization (Bossler & Holt, 2009), and other nonphysical criminal offenses like online fraud (Pratt et al., 2010). However, most of this research has either drawn on smaller samples of self-reported surveys from nonrepresentative data sets (e.g., student surveys), or has focused entirely on a single type of victimization such as cyberbullying or cyberstalking (Leukfeldt & Yar, 2016). We extend this line of research by looking beyond self-reported surveys to study observed online behaviors by analyzing a comparatively larger dataset of Twitter conversations over an extended period (1 month). To understand causes and consequences of online VAW in India, we will follow the three core concepts of RAT: victims, offenders, and guardians. To examine the first concept, we ask:

Research Question 2 (Victims): Do different Indian women of influence receive different types of online harassment on Twitter?

By posing this question, our intention is not to engage in victim blaming or shaming of any kind. Rather we recognize that women are often caught in a career catch-22, where the same personal attributes that increase one's professional success may also increase the chance of being an "attractive" target for harassment. It has been well-documented that many women leaders confront greater prejudice and backlash effects for counterstereotypical behavior in workplace settings (Rudman & Phelan, 2008). We can expect similar behavioral patterns to manifest in online public spheres like Twitter which has historically taken a hands-off approach to content moderation such that problematic content that expresses harm, yet is considered legal, can easily bypass current automated and human-led methods of protective guardianship.

Proponents of RAT surmise that when people engage in more direct forms of online communication, they increase their chance of confronting harmful and threatening behaviors (Leukfeldt & Yar, 2016). The more time an individual spends on social media, the greater chances they will be exposed to hateful material (Costello et al., 2016). Previous research also shows that people who disclose personal information online are more likely to be attacked irrespective of their suitability and "attractiveness" as a target (Welsh & Lavoie, 2012). Holt and Bossler (2008) drew from a self-reported college student survey and showed that daily computer use did not inherently increase the risk of being targeted. Rather, engaging in risky online leisure activities (e.g., illegal downloads) or with users connected to communities of deviance (e.g., pornography, hacking, trolling) exposes one to greater danger (Choi & Lee, 2017). Following this line of examination, we consider characteristics and motivations of offending social media user accounts by asking:

Research Question 3 (Offenders): Who are the posters of online harassment, abuse, and VAW?

Research shows that trolls commonly employ three interrelated strategies: intimidation, shaming, and discrediting when attempting to limit and constrain women's visibility online (Sobieraj, 2018). Other scholars clarify *gendertrolling* as a particularly harmful type of trolling behavior, where misogynists engage in aggressive fear-based tactics (e.g., stalking, threatening insults) in online environments to deter women's visibility and participation (Herring et al., 2002). The motivations behind gendered online offenses are most clearly observed in male-dominated online communities and toxic subcultures, including the #gamergate controversy (Braithwaite, 2016; Massanari, 2017). In the current political climate, there is growing concern over how posters of online VAW might also be more likely to engage in dangerous discourse of othering, hate speech, and extremist content (Awan, 2014).

Our final question focuses on social media guardianship (i.e., community guidelines, automated and human-led content blocking, filters, moderation techniques, user reports), with a particular focus on the effectiveness of Twitter's response to remove harmful content:

Research Question 4 (Guardians): What is the effectiveness of platform-based guardianship?

"Capable" guardianship is the most central and underdeveloped component of the RAT when assessing victimization risks online (Reyns et al., 2016). Physical applications of RAT most commonly conceptualize "capable" guardianship as the capacity (of a person or thing) to effectively protect victims from being targeted, by preventing offenses from occurring. The boundaries of what may be considered a "capable" guardian can vary. For example, guardians can be purposeful and formal (e.g., police); and can also be informal and unintended (e.g., neighbors or coworkers). However,

mechanisms of guardianship are not always capable of protecting targets from motivated offenders (e.g., CCTV).

Compared with victims and offenders, we can start to see why guardianship continues to be the least straightforward in its online translations (Näsi et al., 2017). Just like door locks, car alarms, and home security systems, there are many ways to conceptualize what "capable" guardians[3] resemble online. For this study, we understand platform-based guardianship on social media as automated and human-led content moderation tools (e.g., block, mute, and reporting filters, and community guidelines) that have the capabilities of protecting users from online VAW.

Previous research of "capable" guardianship in online environments has most often focused on software and blocking filters (Mesch, 2009), with varying degrees of success. Hutchings and Hayes (2008) interviewed a random sample of 104 participants and found that email filters were largely ineffective in preventing online phishing attacks. Jansen and Leukfeldt (2016) conducted 30 interviews with victims of online fraud and discovered that even with the help of protective security software, negligence, and lack of digital literacy made certain users more attractive targets for offenders. Navarro and Jasinski (2012) came to a different conclusion in their analysis of safeguards against teenage cyberbullying, showing that online filters significantly decrease and prevent risks of victimization.

The problem with software, blocking, and filter-based forms of guardianship is that these strategies do not translate well to Twitter where users observe and engage in active conversations. Twitter's "shared material architecture" vis-a-vis tagging and retweeting creates an open environment for bystanders to easily become involved in "confrontational encounters" (Udupa, 2018, p. 1512). This affordance structure of Twitter (e.g., attract followers, more retweets, and likes, and) produces incentives for users to express contentious opinions that provoke and spark controversy, in order to gain popularity.

At the same time, user-centered guardianship techniques have been found to be effective because they help moderate content that others will likely find offensive, and public removal promotes positive online community behavior (Gillespie, 2017). Self-moderation of harmful content is almost always done retroactively, and these reactive techniques do not prevent experiences of online victimization from trickling offline. For instance, Jhaver et al. (2018) interviewed a group of Twitter users, and found that despite using blocklists, many people still felt insufficiently protected from online harassment in their daily lives. Considering these findings, we seek to examine Twitter's role and effectiveness in platform-wide guardianship efforts as opposed to individual's efforts to block certain types of content and users.

Methodology

Sample of Indian Women of Influence on Twitter

We began compiling our sample of influential women with all women in lower and upper houses in parliament[4], and expanded our sample to be more representative of

contemporary India. We define influential women as women with key personal attributes (e.g., high credibility, public visibility, specialized expertise) that allow them to motivate change in various realms of contemporary society beyond politics, such as cultural and economic shapers (Bakshy et al., 2011). To enhance diversity, the list of women politicians was then supplemented with a snowball sample of Indian women who have been publicly celebrated as influential leaders (e.g., researchers, writers, journalists, actors, activists, and business moguls) at the time of our data collection.[5] Importantly, we did not categorize or control for the composition of "influence" in our case study because this process would require in-depth research into the Indian context on the ground, which is out of the scope of the current online focused study.

In total, we compiled a sample of 101 influential women and divided our sample into four groups: (a) 59 elected politicians and civil servants with a Twitter account, including 43 out of 65 members of Lok Sabha (Lower House), 11 out of 27 members of Rajya Sabha (Upper House), and 5 civil servants; (b) 16 celebrities from the Indian film and media industry; (c) 12 business women, including executive leaders, entrepreneurs, and CEOs; and (d) 14 other public figures, including 8 activists, 3 journalists, 2 writers, and 1 athlete.

Data Collection

Twitter was selected as the empirical site because it allows public figures to cast a wider net when attracting new audiences and potential followers (Malhotra & Malhotra, 2016). We used Netlytic, an online program for social media text and network analysis to collect and analyze our data. Specifically, we automatically captured any publicly available tweets (original posts or replies) mentioning at least one of Twitter users from our list of 101 accounts. Twitter data were collected over a 1-month period from November 1st to 30th, 2017 (our study period) using Twitter's Search API. We collected a total of 931,363 Twitter messages (excluding duplicates and retweets, referred to as "RTs"). Since our focus is on studying direct and personalized attacks on Twitter, we further removed tweets that can be considered as "mass replies." This is when at least the first 140 characters of a tweet contained a list of user handles. At the end of this cleaning process, we ended up with 720,406 tweets.

Online VAW Dictionaries

For our data-driven research, we used swear words as linguistic cues to detect anger, aggression, and hostility being expressed toward the women in our study. Swearing and offensive hateful commenting behaviors are argued to be driven by the disinhibited and anonymous nature of social media (Cho & Kwon, 2015). It has been well-established that profanity-laced commenting online is emotionally contagious and has spillover effects such as inciting incivility, hostility, and aggressive behaviors between online users (Kwon & Gruzd, 2017; Mead, 2014; Song et al., 2020). In this context, we consider swearing behavior to be indicative of a high arousal of negative and hostile emotions; which can be manifested through direct personal attacks or vitriolic

expressions of disagreement based on a common (unfavorable) characteristic (e.g., career/personal choices, public agendas, specific policies, legislative processes).

To detect swear words, we relied on two dictionaries: South Asian (mostly in Hindi) and English swear words. The South Asian dictionary was iteratively developed by our research team through publicly available sources, including crowdsourced lists of Hindi language swear words on websites such as www.hindilearner.com and www.youswear.com. During this iterative process, we remained mindful that not all insults or abusive comments would be explicitly offensive or blatant swear words. Disrespectful comments with covert, subtle, sarcastic, and insidious types of insults can carry similar weight, contextual meaning, and can have equally harmful impacts (Chaudhry & Gruzd, 2020). Mindful of this, the dictionary was subsequently expanded by manually reviewing over 5,000 tweets by one of the team members who is a Hindi speaker. The resulting South Asian dictionary includes 697 words, including derivatives, abbreviations, and slang. The English dictionary consisted of 580 items (keywords and phrases) that was created based on Kwon and Gruzd (2017).

Two different coders (one postdoc and one graduate research assistant) from our team undertook a manual content analysis of sample offensive tweets to validate the swearing dictionaries. Each coder manually reviewed 1,200 tweets from each user group, as flagged by the two dictionaries. A total of 4,800 tweets were analyzed for each swearing dictionary. Words and phrases generating more than 10% of false positives were removed from each dictionary. This was a complex task given that we were working with two primary languages, where words could have two different meanings depending on the language. For example, the English phrase *pooch* is often understood to mean dog, which can be used as a derogatory slur toward women, whereas in Hindi, the (English-alphabetized) phrase *pooch* could translate to "ask" or to "inquire." Once developed and refined, both dictionaries were used to automatically detect any tweets with swear words that may be indicative of an overt form of online VAW.

Data Analysis

The analysis used a mixed-methods approach, which combines results from an automated text analysis to detect explicit harassment, a manual examination of sample tweets to explore nuances in the covert messages and implicit forms of harassment, and a qualitative review of public user profiles among the most active online VAW offenders.

Specifically, to assess the prevalence of swearing tweets (Research Question 1), we used a custom R script to automatically identify tweets that contained one or more items (words or phrases) from our dictionaries of swear words. Next, to examine the types of VAW and the types of offenders based on the four user groups (Research Question 2 and Research Question 3), we undertook a manual review of the types of posts and public profiles of frequent offenders, as automatically flagged by our dictionaries. Finally, to determine how well Twitter deals with detecting and removing VAW-type posts (Research Question 4), we used a program called Hydrator[6] and a

Python library called Twarc to check which tweets from our initial data set are still accessible a year after the original data collection in November 2017 or if any of the tweets/accounts was removed by a user or blocked by the platform.

Results

Research Question 1

Instances of online VAW were most often situated in response to current affairs being discussed in popular Indian news media outlets, which at the time of our data collection focused on the tense relations between the ruling BJP, their Hindu nationalist supporters, left-wing secular activists, and minority groups (Menon, 2018). Many journalists have voiced concern over the public intimidation and backlash they receive from Hindu nationalists when criticizing PM Modi and BJP government policies (Gopalakrishnan, 2018). Our results confirm these reports, as we found that many women of influence, who are outspoken in their opinions, similarly confronted offensive tweets questioning their group loyalty to India.

Through the qualitative analysis of sample tweets, we also discovered that South Asian and English dictionaries shared thematic similarities in the type of words and phrases expressed in offensive tweets. Our examination of sample offensive tweets reveals online VAW to fall into three broad categories: *dismissive insults, ethnoreligious slurs*, and *gendered sexual harassment.* Table 1 shows frequently used swear words grouped by these three broad categories. While there are similarities in the types of offensive words and phrases, instances of gender-based harassment flagged by the South Asian dictionary were more sexually explicit.

The percentage of tweets that contained one or more swear words/phrases as flagged by our dictionaries ranged from 0.4% to 4.8% per group (see Figure 1). The highest percentage was for Group 4 (“Other Public Figures”) based on the English dictionary, and the lowest was for Group 3 (“Business”) based on the South Asian dictionary. Our result is consistent with previous work that finds swearing, dismissive insults, and abusive words to make up around 3% of online communications (Subrahmanyam et al., 2006). Although the overall percentage may seem low; even a single harassing message can be threatening to an individual, especially those seeking to enter public life for which an active social media presence is the norm. Importantly, the percentages do not include any harassing messages that were retweeted (reposted) as we only focused on original and reply messages for this study. Furthermore, because the Twitter API truncated long RTs and replies to 140 characters at the time of our data collection, only tweets that mentioned one of the seed accounts in the first 140 characters of a post were used for this analysis. Thus, considering that some truncated tweets might have included swearing that was missing in our data, and also because our dictionaries only flagged explicit forms of swearing, we expect the total percentage of harassing tweets might be higher. But even keeping in mind some of our methodological limitations, the result shows that different groups of influential women are being attacked at a different rate as discussed in detail under Research Question 2 next.

Table 1. Types of 'Swearing' Tweets.

Broad types	Frequently used swear words and phrases	
	South Asian dictionary	English dictionary
Dismissive insults	*pagal* = stupid/crazy; *chamcha* = ass kisser/suckup; *kamini* = scoundrel; *deshdrohi* = traitor; *chutiya* = fucker; *chor* = thief; *madarchod* or "MC" = motherfucker; *bahenchod* or "BC" = sister fucker; *janwar* = animal; *bakra* = goat; *chup ho* = shut up	stupid; retard; idiot; dumb; greedy; coward; sickular; libtard; asshole; fuck u; anti-Hindu, antinational; scum; parasite; bootlicker
Ethnic slurs	*jihadi aurat* = terrorist woman; *saur* = pig; *mulli, porki, porkistani* = derogatory term for Muslim	shameless fascist jihadi thug; go fuck yourself you Pakistani bitch; Islamist hypocrite; commy pig; terrorist; jihadi; Muslim lunatic
Gendered and sexualized harassment	"presstitute"; *mujra* or *randi* = hooker/prostitute; *lund choos* = suck dick; TMC or *teri ma ki chut* = you mother's vagina; *kutti* = dog/bitch; *gandi aurat* = dirty woman; *buddiya* = old lady; *moti* = fat	bitch; witch; bollywood whore; slut; cultureless cunts psycho auntie; disgusting homewrecker; shagged senseless; fuck her hard; ugly piece of shit

Research Question 2

Our manual review of flagged tweets suggests that different women of influence (depending on their career choices) received different types of online harassment, abuse, and expressions of violence. The majority of offensive messages received by *Group 1: Politicians* (comprising 1.7% and 2.2% of messages based on the South Asian and English dictionaries correspondingly) are not directed to a single person. These messages were found to often engage in dismissive and expletive insults that are sometimes accompanied with Islamophobic ethnoreligious slurs. Women politicians commonly receive offensive tweets that are reactionary and express discontent based on policies, party agendas, and public statements made by governing officials. For instance, "You must either be a total retard or must hate India @username . . . Ajeebbjpgang."

While further statistical analysis is required, our manual review of tweets showed that different political stances (e.g., Congress vs. BJP party) did not play a decisive role in the types of vitriol directed to our sample of women politicians. In addition to explicit swearing behavior, many offensive tweets directed to this group attempt to dismiss the legitimacy of women politicians based on intellectual ability and patriotic commitments to India. For example, "@username Ohh shut up You pro Islamist raving and ranting good for nothing You are a disgrace to this land."

By contrast, offensive messages received by *Group 2: Celebrities* (comprising 1.5% and 2.3% of messages based on the South Asian and English dictionaries correspondingly) were far more gendered. Messages included explicit swearing, sexual harassment, and engaged in slut-shaming, and fat-shaming behaviors. Here is a sample tweet in reply to one of the celebrity accounts from this group: "@username A whore is better than you." For this group, we found offensive South Asian tweets

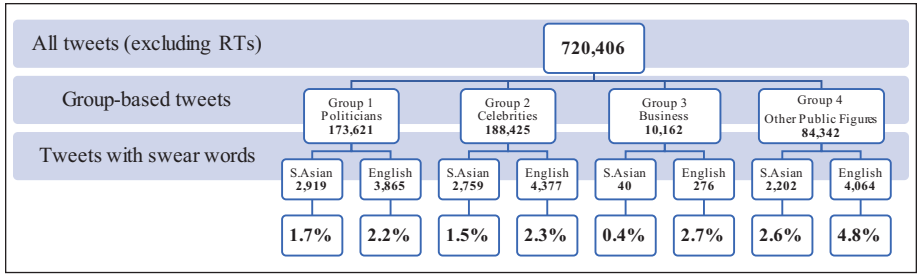

Figure 1. Percentages of swearing tweets by user group and by dictionary.
Note. RTs = retweets.

often employ context-specific Indian *Gaali* slangs that combine humor with sarcastic insults and abuses that are mostly used in informal conversations between everyday people such as the following: "May be @username got new implants and need some attention . . . Randi saali[7] #Entertainment."

Offensive tweets received by *Group 3: Business* (comprising 0.4% and 2.7% of messages based on the South Asian and English dictionaries correspondingly) demonstrated more critical commentary based on investment and entrepreneurial ventures. These offensive tweets are found to express dismissive insults that question intellectual competence, qualifications, and/or new public-private partnerships; for example, "I am smart, i can track an idiot @username,who som[e]how managed to be a CEO & spoiling a name of #India !!." Our manual review of sample tweets also reveals thematic similarities between messages flagged by the South Asian and English dictionaries. For Group 3, both swearing dictionaries uncovered significantly fewer instances of ethno-religious forms of hateful speech. We found that Group 3 received fewer offensive tweets overall compared with the other groups analyzed for this study.

Finally, offensive tweets received by *Group 4: Other Public Figures* such as journalists, writers, activists, and so on (comprising 2.6% and 4.8% of messages based on each dictionary), expressed more direct forms of gendered and ethnoreligious online harassment, including death threats in rare cases. Tweets like "when 1000 pigs 🐖 died you born. Don't deserve to be in this country" and "What an idiot u r . . . shameless fellow . . . u should be kicked u moron" were common for this group.

This is not surprising as many of these women use social media to openly comment and share their opinions on public issues. Not only do many of these women receive blatant forms of online sexual harassment, they also commonly confront Islamophobic slurs and are dismissed as antinationals based on their political views and commentary. Overall, we found that for Group 4 the recorded offensive tweets are far more violent and explicit than previous groups. Future work ought to confirm this result with statistical testing of cross-group differences in terms of the prevalence and the types of messages received by women in each group.

Research Question 3

For this portion of our analysis, we conducted a manual examination of the most active Twitter users[8] posting offensive swearing tweets (based on the number of detected posts). Specifically, we reviewed 80 public profiles and their recent tweets for the top 10 posters of online harassment for each of the four groups and based on each of the South Asian and English swearing dictionaries. We discovered that offenders engaging in attacks are more likely to be individually motivated than to be part of an organized campaign effort against a single person or account. Many offenders engage in persistent trolling behaviors, with most of their recent tweets expressing negative and directing harmful sentiments towards specific women. Our manual review of these accounts also suggests that offender accounts can be grouped into three broad categories: *News Junkies, Bollywood Fanatics*, and *Lone-Wolves*.

News Junkies' are Twitter users interested in the latest Indian breaking news, sports, cultural, and political affairs. Posts from these accounts look to be from individuals located in India, many of whom appear to be sympathetic to a political cause or movement (i.e., BJP supporters, self-proclaimed PM Modi fans, Hindu nationalists, proud Indian secularists). This group of users differed in their Twitter following (ranging from 26 followers to under 5,000) and number of followers (ranging from 20 to over 800), and rarely engage in targeted attacks toward individual accounts.

Bollywood Fanatics describe users particularly interested in Indian pop culture and entertainment. Posts from these user accounts often focused on popular Bollywood actors. Many of the flagged tweets for this category of offenders are in conversation with other Twitter accounts including prominent celebrities, news media outlets, and journalists. They also delve into celebrity gossip and updates on the latest movie releases, television shows, award shows, and fashion choices, which explains why body-slut shaming tweets were flagged by our swearing dictionaries.

Lone-Wolves are Twitter accounts that purposefully antagonize and troll women of influence. This group of offenders exhibit varying negative posting behaviors that require further unpacking. In addition, because the present study did not test a control group, it remains unclear whether these "lone-wolf" accounts would be likely to express violence differently against other gender identities. For example, we found some offenders to be overtly antagonistic and persistently aggressive throughout many of their postings which could spark more offensive engagement behaviors in other Twitter users. We also found there to be Twitter users (with comparatively less followers to following) who did not interact frequently with other users, yet consistently engaged in direct online harassment against individual accounts from our sample. This could indicate bot-like behaviors that would require further analysis.

Research Question 4

Based on our analysis (see Table 2), as of April 17, 2019, the majority of swearing posts flagged by our two dictionaries have not been deleted. About 5% to 13% of the flagged tweets are no longer available because they were deleted by the original

Table 2. Deletion Rate of "Swearing" Tweets.

	Group 1: Politicians		Group 2: Celebrities		Group 3: Business		Group 4: Other public figures	
Tweet status	SA	ENG	SA	ENG	SA	ENG	SA	ENG
TWEET_OK	77%	78%	74%	74%	83%	77%	72%	74%
TWEET_DELETED	10%	9%	12%	12%	5%	13%	11%	11%
USER_PROTECTED	3%	2%	2%	2%	n/a	2%	1%	2%
USER_SUSPENDED	11%	12%	12%	12%	13%	8%	16%	14%
Total Tweets	2,919	3,865	2,759	4,377	40	276	2,202	4,064

Note. SA = South Asian dictionary; ENG = English dictionary.

poster or because the account was deleted by the user themselves. 1% to 3% of the flagged tweets are no longer available publicly because the original poster has changed their privacy setting of their account to "protected." Only 8% to 16% of the flagged tweets were removed because Twitter has suspended the original poster, with the highest rate for tweets targeting our sample users from Group 4. Our expectation was that this percentage would be much higher considering the explicit nature of swearing and harassment in the detected posts. The following may explain why this was not the case.

In 2018, Twitter undertook a massive overhaul of its content moderation algorithm (Wong, 2018) by shifting from content-based to conduct-based guardianship; where the behavioral signals of Twitter's online community are used to determine when a user is detracting from (rather than adding to) the overall tone of a conversation. In this new approach, offensive tweets are pushed down further into a larger list of search results and/or replies that makes them difficult to retrieve or see in the first instance. The problem with this form of guardianship, however, is that some offensive tweets while hidden are not permanently deleted from the platform and thus remain accessible to users. In addition, these algorithmic changes are unlikely to prevent motivated users from posting offensive messages.

Discussion and Conclusions

By using the case of India, the present study sought to explore the prevalence and patterns of online VAW on Twitter. While we recognize that online VAW in India has its own specific challenges, India also represents a microcosm of a dangerous phenomenon taking place and being normalized elsewhere in the world. Ultimately, our study shows that in the social media age, online and offline public spheres intertwine, requiring improved tools of "capable" guardianship that support and empower women to actively participate in public life free and fairly.

Our study contributes to widening of applications of the "cyber–lifestyle" RAT in multiple ways. It shows that social media platforms like Twitter create opportunity

structures for motivated offenders to readily exploit and engage in harmful behavior like online VAW. Social media affordances (e.g., likes, direct messages, RTs) also play important roles in supporting "herding" and "bandwagoning" like behaviors which can be difficult for women to navigate when being targeted online (Ben-David & Matamoros-Fernández, 2016). As such, our research interrogates the concept of "capable" guardianship by showing that content moderation tools afforded by Twitter are not fully effective in protecting and preventing Indian women of influence from receiving offensive tweets.

We argue that RAT's traditionally blunt and heavy-handed framings of guardianship should be enhanced to better fit within and reflect the conditions surrounding our present social media reality. Women's safety and security are real-life challenges, and online–offline spaces can no longer be viewed as distinctly separate in public life. Like bullets to a gun, motivated offenders can use their words as ammunition to harm with low cost and considerable ease on social media. Our research showed that preexisting gender inequalities can easily transpose online. Hence, without improved guardianship capabilities, we can expect that social media will continue to present obstacles of inclusion for women.

The analysis for example reveals that instances of online VAW in the Indian context are not always organized offender campaigns or direct targeted attacks. Instead, most offensive tweets are situated in response to various news media updates, political agendas, government policies, and everyday sociocultural affairs. This might explain why we found motivated offenders (e.g., *news junkies, Bollywood fanatics*, and *lone-wolves*) largely directed their offensive vitriol toward multiple people and/or accounts, that in their eyes shared similar unfavorable characteristics. This finding is important because it suggests that Twitter users who engage in online VAW in the Indian context do so openly and thus have slowly normalized offensive language as part of their online vernacular.

In our study, most of the observed offensive tweets could be categorized as: *dismissive insults, ethnoreligious slurs*, and/or *gendered sexual harassment* (see Table 1). Importantly, the South Asian and English dictionaries used to facilitate our analysis were linguistically different, and yet, shared thematic similarities in the types of harmful words and phrases expressed in the discourse of online VAW. However, instances of gender-based sexual harassment flagged by the South Asian dictionary were far more sexually explicit compared with those flagged by the English dictionary. These findings point to distinct Indian *Gaali* cultures, where offenders intentionally frame their sexually charged vitriol with sarcasm and humor as a way to simultaneously participate, gain traction, and intimidate women in the Indian online public sphere (Udupa, 2018). This finding is important because it shows how cultural context is an equally relevant point of consideration when developing mechanisms of platform-based guardianship around content moderation and management strategies.

We discovered that women journalists, academics, and activists were targeted with more extreme threats of violence. The lack of "capable" guardianship on social media has led numerous women in public roles in other countries to censor their online activities due to troubling experiences with harassment, abuse, and violence (Amnesty International, 2018; Jane, 2016). Problematically, the growing weaponization of

speech coupled with a normalized culture of online misogyny stifles and silences women while simultaneously amplifying their "attractiveness" as targets for harassing, abusive, and violent communications.

From our "cyber–lifestyle" RAT lens, notions of "capable" guardianship on social media should be further developed; keeping in mind that *prevention* and *punishment* of online VAW increasingly requires multilevel strategic responses from diverse sets of actors. This includes considering how formal (platform rules, codes of conduct, user reports, automated content filters) and informal (bystanders, user flagging, blocking, and reporting) methods of moderation may be combined to strengthen guardianship capabilities. Future developments in platform-based methods of guardianship should therefore push for a reevaluation of online VAW as a much broader digital infrastructure systems problem (Eck & Clarke, 2003).

The need for more comprehensive platform-based guardianship becomes clearer when considering the negative political, social, cultural, and health-related consequences associated with online VAW. There is a general consensus among scholars that current regulatory and content moderation approaches are ill-equipped to manage the scale, intensity, and global reach of harmful online content that infringe on the rights and freedoms of vulnerable groups (Common, 2020; Gillespie, 2018; Gorwa et al., 2020).

Underpinning our results, is evidence of a broader need for further cooperative partnerships between governing stakeholders, social media platforms and tech industry leaders, Internet researchers, and global civil society to address the issue of online VAW. As social media becomes further engrained into our lives, so too has the push for greater "cooperative responsibility" beyond a single central actor; with platforms, public institutions, and users increasingly being called on to collaboratively develop public values, regulations, policies, and solutions to dangerous online threats like VAW (Helberger et al., 2018). Thus far, institution-led governing bodies and industry-led social media platforms have been slow to cross-pollinate, share ideas and information that might otherwise help prevent online VAW (Tenove et al., 2018). Comprehensive revisions in social media regulatory frameworks and content moderation policies are essential to ensuring that platforms like Twitter do not inadvertently create hostile environments for women. Next, focusing on the Indian context, we provide a summary of what policies and tools are being considered and implemented by governments, social media platforms, and users.

Government Policies and Initiatives: India

The current Indian legal framework protects victims and punishes offenders of online VAW through two key pieces of legislation: the Indian Penal Code (IPC) and the Informational Technology (IT) Act (Uma, 2017). The Ministry of Electronics and Information Technology (MeitY) has been especially involved in shifting IT guidelines and policy discussions around social media platforms, urging intermediaries to take-action against online content that may affect the public negatively. Problematically, some of these recommendations and laws do not explicitly include terms like online

harassment or violence, and this definitional gap has led to differing interpretations on whether and to what degree online VAW causes harm in real-life. Despite these definitional loopholes, there have been notable landmark cases involving prominent figures (actor Parvathy) that have shown current legislative frameworks can successfully punish and convict offenders of online VAW (Awasthi, 2018).

Indian MP Maneka Gandhi has been especially active in addressing online VAW in recent years. In 2016, Gandhi led a social media campaign against online trolling (Express News Service, 2016). As part of this campaign, women were encouraged to use the hashtag #IamtrolledHelp to log their complaints with the National Commission for Women. Problematically, these types of initiatives require women to publicly declare their experiences of online VAW, which can be traumatic and can have negative social repercussions. In 2017, the Indian government launched the "I am Trolled" smartphone app initiative to encourage Indian women and girls to report and seek help against online violence and physical harassment (Indo-Asian News Service, 2017). However, apps rely on self-reporting *after* the offense has taken place. Using the "cyber–lifestyle" RAT, these retroactive tools of guardianship are incapable of disrupting and preventing "motivated" offenders, nor do they make targets any less "attractive" online.

Social Media Platforms and Guardianship: Twitter

Twitter has long been criticized for the seemingly hands-off approach to online harassment and abuse when defending postings of harmful content (Wagner, 2017). There have been several high-profile cases of women celebrities, journalists, and politicians who have experienced misogynistic violence and endured threatening insults on Twitter which has put added spotlight on Twitter's lack of response to abuse in recent years.

In 2017, a large-scale global protest using the hashtag #WomenBoycottTwitter resulted in CEO Jack Dorsey vowing to crack down on harmful and abusive activity. After facing much public criticism Twitter started working with its Trust and Safety Council, taking a more active stance in harmful content conversations (Flynn, 2017). Prominent strategies for combating online VAW still rely on self-reporting mechanisms, content flagging, and keyword blocking filters, which again places responsibility of "capable" guardianship on victims and/or bystanders (Green, 2016; Twitter, n.d.-b).

Since the completion of this research, Twitter has iteratively widened the platform-guardianship around hateful speech and harmful content for its users to flag dehumanizing language between 2019 and 2020 (Twitter, n.d.-a). As of May 2020, Twitter's Hateful conduct policy focuses on combating dehumanizing and insulting speech that targets a group of people because of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. The context-based expansion of content moderation is considered a step in the right direction because it acknowledges that speech does not have to be blatant or explicitly hateful to cause harm. Importantly, our findings show that online VAW in the Indian context commonly manifests through ethnoreligious othering and Islamophobic slurs that in the current framework would be deemed as dehumanizing.

Limitations and Future Directions

The current study contributes an empirical analysis of online VAW in the Indian context on Twitter. The results presented in the research did not include any harassing messages that were retweeted (reposted), as we only focused on original and reply messages for this study.

In addition, because the study period began in November 2017, the Twitter data collected was limited to 140 characters (Twitter has since doubled the allowance of characters to 280 per post). Furthermore, messages that were flagged as part of our analysis included explicit forms of swearing (based on our two dictionaries). Implicit, covert, and subtle forms of abusive language and image-based forms of online harassment were not included in the current analysis. Future work should also consider how temporality (election cycles) may affect instances of online VAW. Another potential fruitful area of research could compare instances of online VAW, and its spread across different social media platforms (e.g., Facebook, Instagram, Reddit).

Research should also focus on the offenders of online VAW to gain a more nuanced understanding of what motivates people to target women on social media. In addition, because our study did not have a control group to test and compare our findings, an avenue for further work should be to investigate the different types of online harassment women of influence confront compared to other gender identities. Moving beyond the Indian context, it is recommended that future work expand the global scale of this research by delving into other regions, countries, geopolitical, cultural, and linguistic contexts.

Acknowledgments

We would like to thank members of the Ryerson University Social Media Lab for providing feedback throughout the project and especially to Jenna Jacobson, Michael Pacheco, Amrita Maharaj, Nadia Conroy, Nikolai Krause and Lilach Dahoah Halevi. We also thank anonymous reviewers and the special issue editors, Hazel Kwon, Weiai Xu, and Barry Wellman for providing very helpful comments.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research is supported in part by funding from Global Affairs Canada and Social Sciences and Humanities Research Council of Canada (PI: Gruzd).

Notes

1. The 2014 #Gamergate controversy targeted women from the videogame industry with online harassment, threats of rape, and abusive comments that were so severe that some women, such as Zoe Quinn and Anita Sarkeesian, went into protective hiding for their

own safety. "The Fappening" or "Celebgate" took place in 2014, when hundreds of private images of celebrity actors were shared and leaked (otherwise called "doxing") on 4Chan, Reddit, and Tumblr.

2. If compared with other countries, India has the 2nd highest percentage of Twitter visitors (after the United States).
3. This includes social media platforms, moderation styles, social media users, and other platform affordances that have the capacity to reduce and/or prevent online VAW by protecting targeted users.
4. See PRS Legislative Research (http://web.archive.org/web/20210105115411/https://www.prsindia.org/).
5. Sources include the following: *India Today*, "India's 25 most influential women," March 2013; *India TV News*, "India's 110 most powerful female politicians," May 2014; *Aapka Times* "10 Most influential women student leaders of India," April 2016; *Economic Times*, "The 20 most influential global Indian women," January 2015; *India.com* "International Women's Day 2017: Top 8 Women Leaders from India who influence people worldwide," March 2017; *Youth Ki Awaaz*, "Top 10 women entrepreneurs and leaders of India," January 2011.
6. Hydrator app (see https://github.com/DocNow/hydrator).
7. The phrase "Randi saali" broadly translates to prostitute in this context.
8. Their user handles/names are not included in this study for privacy considerations.

References

Alexa.com. (2021, January 20). *Twitter.com Competitive Analysis, Marketing Mix and Traffic*. https://www.alexa.com/siteinfo/twitter.com#section_traffic

Amnesty International. (2017). *Amnesty reveals alarming impact of online abuse against women*. https://www.amnesty.org/en/latest/news/2017/11/amnesty-reveals-alarming-impact-of-online-abuse-against-women/

Amnesty International. (2018). *Toxic Twitter—The silencing effect*. https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-5/

Amnesty International India. (2018). *Why we need to talk about online violence against women in India*. https://amnesty.org.in/need-talk-online-violence-women-india-2/

Awan, I. (2014). Islamophobia and Twitter: A typology of online hate against Muslims on social media. *Policy & Internet*, *6*(2), 133-150. https://doi.org/10.1002/1944-2866.POI364

Awasthi, S. (2018, January 2). Online trolls beware, these are the laws that could be used against you. *The Indian Express*. http://indianexpress.com/article/india/online-trolls-beware-these-are-the-laws-that-could-be-used-against-you-5008637/

Bakshy, E., M., Hofman, J., Mason, W., & Watts, D. (2011, February). Everyone's an influencer: Quantifying influence on Twitter. In *Proceedings of the fourth ACM International Conference on Web Search and Data Mining* (pp. 65-74). https://doi.org/10.1145/1935826.1935845

Ben-David, A., & Fernández, A. M. (2016). Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-Right political parties in Spain. *International Journal of Communication*, *10*, 1167-1193.

Bhargava, Y. (2017, October 5). 8 Out of 10 Indians have faced online harassment. *The Hindu*. http://www.thehindu.com/news/national/8-out-of-10-indians-have-faced-online-harassment/article19798215.ece

Bossler, A. M., & Holt, T. J. (2009). On-line activities, guardianship, and malware infection: An examination of routine activities theory. *International Journal of Cyber Criminology*, *3*(1), 400-420. http://www.cybercrimejournal.com/bosslerholtjan2009.htm

Braithwaite, A. (2016). It's about ethics in games journalism? Gamergaters and geek masculinity. *Social Media + Society*, *2*(4), Article 667248. https://doi.org/10.1177/2056305116672484

Burgess, J., & Matamoros-Fernández, A. (2016). Mapping sociocultural controversies across digital media platforms: One week of #gamergate on Twitter, YouTube, and Tumblr. *Communication Research and Practice*, *2*(1), 79-96. https://doi.org/10.1080/22041451.2016.1155338

Chachra, D. (2017). To reduce gender biases, acknowledge them. *Nature News*, *548*(7668), Article 373. https://doi.org/10.1038/548373a

Chaturvedi, A. (2017, May 17). How India emerged as Twitter's fastest growing market in terms of daily active users. *Economic Times*. https://economictimes.indiatimes.com/opinion/interviews/india-became-our-number-one-market-in-daily-users-twitters-new-india-director-taranjeet-singh/articleshow/58601906.cms

Chaudhry, I., & Gruzd, A. (2020). Expressing and challenging racist discourse on Facebook: How social media weaken the "spiral of silence" theory. *Policy & Internet*, *12*(1), 88-108. https://doi.org/10.1002/poi3.197

Chess, S., & Shaw, A. (2015). A conspiracy of fishes, or, how we learned to stop worrying about #gamergate and embrace hegemonic masculinity. *Journal of Broadcasting & Electronic Media*, *59*(1), 208-220. https://doi.org/10.1080/08838151.2014.999917

Cho, D., & Kwon, K. H. (2015). The impacts of identity verification and disclosure of social cues on flaming in online user comments. *Computers in Human Behavior*, *51*(Part A), 363-372. https://doi.org/10.1016/j.chb.2015.04.046

Choi, K.-S., & Lee, J. R. (2017). Theoretical analysis of cyber-interpersonal violence victimization and offending using cyber-routine activities theory. *Computers in Human Behavior*, *73*(August), 394-402. https://doi.org/10.1016/j.chb.2017.03.061

Cohen, L., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review*, *44*(4), 588-608. https://doi.org/10.2307/2094589

Common, M. F. (2020). Fear the reaper: How content moderation rules are enforced on social media. *International Review of Law, Computers & Technology*, *34*(2), 126-152. https://doi.org/10.1080/13600869.2020.1733762

Cooney, S. (2018, March 6). Bumble's CEO, on why the dating app is banning photos with guns. *Time*. http://time.com/5188598/bumble-gun-ban-ceo-nra/

Correa, T., & Pavez, I. (2016). Digital inclusion in rural areas: A qualitative exploration of challenges faced by people from isolated communities. *Journal of Computer-Mediated Communication*, *21*(3), 247-263. https://doi.org/10.1111/jcc4.12154

Costello, M., Hawdon, J., Ratliff, T., & Grantham, T. (2016). Who views online extremism? Individual attributes leading to exposure. *Computers in Human Behavior*, *63*(October), 311-320. https://doi.org/10.1016/j.chb.2016.05.033

Darade, P. (2017, October 7). All you need to know about PM's Pradhan Mantri Gramin Digital Saksharta Abhiyan. *India.Com*. http://www.india.com/news/india/narendra-modi-to-launch-pradhan-mantri-gramin-digital-saksharta-abhiyan-heres-everything-you-need-to-know-2519187/

Dutt, B. (2017, May 12). Let's talk about trolls: Online abuse a weapon to silence women. *Hindustan Times*. https://www.hindustantimes.com/india-news/let-s-talk-about-trolls-trolling-is-a-weapon-to-silence-women-barkha-dutt/story-A9X3fAuRwZiwVrhYQnK-bYL.html

Express News Service. (2016, July 6). Maneka Gandhi takes on cyber trolls, asks women victims to inform her. *The Indian Express*. http://indianexpress.com/article/india/india-news-india/maneka-gandhi-takes-on-cyber-trolls-asks-women-victims-to-inform-her-2896042/

Felson, M., & Eckert, M. (2015). *Crime and everyday life*. Sage.

Flynn, K. (2017, December 18). The "Twitter purge" Nazi reckoning has begun: Here are the rules. *Mashable*. https://mashable.com/2017/12/18/twitter-purge-neo-nazi-reckoning-new-rules-hate-speech/

Fuchs, C. (2014). Social media and the public sphere. *TripleC: Communication, Capitalism & Critique*, *12*(1), 57-101. https://doi.org/10.31269/triplec.v12i1.552

Gillespie, T. (2017). Governance of and by platforms. In J. Burgess, A. Marwick, & T. Poell (Eds.), *The SAGE handbook of social media* (pp. 254-278). Sage.

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Ging, D., & Siapera, E. (2018). Special issue on online misogyny. *Feminist Media Studies*, *18*(4), 515-524. https://doi.org/10.1080/14680777.2018.1447345

Gopalakrishnan, R. (2018, April 27). Indian journalists say they intimidated, ostracized if they criticize Modi and the BJP. *Reuters*. https://www.reuters.com/article/us-india-politics-media-analysis/indian-journalists-say-they-intimidated-ostracized-if-they-criticize-modi-and-the-bjp-idUSKBN1HX1F4

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), Article 897945. https://doi.org/10.1177/2053951719897945

Gray, K. L. (2012). Intersecting oppressions and online communities. *Information, Communication & Society*, *15*(3), 411-428. https://doi.org/10.1080/1369118X.2011.642401

Green, E. (2016, August 18). Why blocking trolls doesn't work. *Time*. http://time.com/4457275/twitter-blocking-troll-failure/

Gruzd, A., & Haythornthwaite, C. (2013). Enabling community through social media. *Journal of Medical Internet Research*, *15*(10), e248. https://doi.org/10.2196/jmir.2796

Habermas, J. (1989). *On society and politics: A reader*. Beacon Press.

Helberger, N., Pierson, J., & Poell, T. (2018). Governing online platforms: From contested to cooperative responsibility. *Information Society*, *34*(1), 1-14. https://doi.org/10.1080/01972243.2017.1391913

Herring, S., Job-Sluder, K., Scheckler, R., & Barab, S. (2002). Searching for safety online: Managing "trolling" in a feminist forum. *Information Society*, *18*(5), 371-384. https://doi.org/10.1080/01972240290108186

Holt, T. J., & Bossler, A. M. (2008). Examining the applicability of lifestyle-routine activities theory for cybercrime victimization. *Deviant Behavior*, *30*(1), 1-25. https://doi.org/10.1080/01639620701876577

Hutchings, A., & Hayes, H. (2008). Routine activity theory and phishing victimisation: Who gets caught in the net. *Current Issues in Criminal Justice*, *20*(3), 433-452. https://doi.org/10.1080/10345329.2009.12035821

Indo-Asian News Service. (2017, March 3). Government to launch anti-troll app for women. *Hindustan Times*. https://www.hindustantimes.com/india-news/government-to-launch-anti-troll-app-for-women/story-EKRi2JnS0j80dbC80sBZFK.html

Jane, E. A. (2016). Online misogyny and feminist digilantism. *Continuum*, *30*(3), 284-297. https://doi.org/10.1080/10304312.2016.1166560

Jansen, J., & Leukfeldt, R. (2016). Phishing and malware attacks on online banking customers in the Netherlands: A qualitative analysis of factors leading to victimization. *International Journal of Cyber Criminology*, *10*(1), 79-91. https://doi.org/10.5281/zenodo.58523

Jhaver, S., Ghoshal, S., Bruckman, A., & Gilbert, E. (2018). Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction*, *25*(2), Article 12. https://doi.org/10.1145/3185593

Kwak, H., Blackburn, J., & Han, S. (2015). *Exploring cyberbullying and other toxic behavior in team competition online games*. http://arxiv.org/abs/1504.02305

Kwon, K. H., & Gruzd, A. (2017). Is offensive commenting contagious online? Examining public vs interpersonal swearing in response to Donald Trump's YouTube campaign videos. *Internet Research*, *27*(4), 991-1010. https://doi.org/10.1108/IntR-02-2017-0072

Leukfeldt, E. R., & Yar, M. (2016). Applying routine activity theory to cybercrime: A theoretical and empirical analysis. *Deviant Behavior*, *37*(3), 263-280. https://doi.org/10.1080/01639625.2015.1012409

Malhotra, C. K., & Malhotra, A. (2016). How CEOs can leverage Twitter. *MIT Sloan Management Review*, *57*(2), 72-79. https://sloanreview.mit.edu/wp-content/uploads/2015/12/298f9b44201.pdf

Marwick, A. (2013, March 29). Donglegate: Why the tech community hates feminists. *WIRED*. https://www.wired.com/2013/03/richards-affair-and-misogyny-in-tech/

Massanari, A. (2017). #Gamergate and the fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, *19*(3), 329-346. https://doi.org/10.1177/1461444815608807

Mead, D. (2014, February 19). People sure tweet "fuck" a lot, finds science. *Motherboard*. https://motherboard.vice.com/en_us/article/8qxn8a/people-sure-tweet-fuck-a-lot-says-science

Menon, A. (2018, April 26). Left turn ahead: CPM-Congress alliance to keep the BJP in check? *India Today*. https://www.indiatoday.in/magazine/the-big-story/story/20180507-congress-cpim-alliance-to-beat-bjp-general-elections-2019-1221666-2018-04-26

Mesch, G. S. (2009). Parental mediation, online activities, and cyberbullying. *CyberPsychology & Behavior*, *12*(4), 387-393. https://doi.org/10.1089/cpb.2009.0068

Mitter, S. (2015, January 19). How Twitter changed its mind on India. *Forbes India*. http://www.forbesindia.com/article/big-bet/how-twitter-changed-its-mind-on-india/39391/1

Näsi, M., Räsänen, P., Kaakinen, M., Keipi, T., & Oksanen, A. (2017). Do routine activities help predict young adults' online harassment: A multi-nation study. *Criminology & Criminal Justice*, *17*(4), 418-432. https://doi.org/10.1177/1748895816679866

Navarro, J. N., & Jasinski, J. L. (2012). Going cyber: Using routine activities theory to predict cyberbullying experiences. *Sociological Spectrum*, *32*(1), 81-94. https://doi.org/10.1080/02732173.2012.628560

Neal, S. (2017). *The surprising reason why CEOs should be social media savvy*. https://www.cnbc.com/2017/04/13/the-surprising-reason-why-ceos-should-be-social-media-savvy.html

Pratt, T., Holtfreter, K., & Reisig, M. (2010). Routine online activity and internet fraud targeting: Extending the generality of routine activity theory. *Journal of Research in Crime and Delinquency*, *47*(3), 267-296. https://doi.org/10.1177/0022427810365903

Reyns, B. W., Henson, B., & Fisher, B. (2011). Being pursued online: Applying cyberlifestyle–routine activities theory to cyberstalking victimization. *Criminal Justice and Behavior*, *38*(11), 1149-1169. https://doi.org/10.1177/0093854811421448

Reyns, B. W., Henson, B., & Fisher, B. S. (2016). Guardians of the cyber galaxy. *Journal of Contemporary Criminal Justice*, *32*(2), 148-168. https://doi.org/10.1177/1043986215621378

Rudman, L. A., & Phelan, J. E. (2008). Backlash effects for disconfirming gender stereotypes in organizations. *Research in Organizational Behavior*, *28*, 61-79. https://doi.org/10.1016/j.riob.2008.04.003

Shaw, A. (2014). The internet is full of jerks, because the world is full of jerks: What feminist theory teaches us about the internet. *Communication and Critical/Cultural Studies*, *11*(3), 273-277. https://doi.org/10.1080/14791420.2014.926245

Sobieraj, S. (2018). Bitch, slut, skank, cunt: Patterned resistance to women's visibility in digital publics. *Information, Communication & Society*, *21*(11), 1700-1714. https://doi.org/10.1080/1369118X.2017.1348535

Song, Y., Kwon, K. H., Xu, J., Huang, X., & Li, S. (2020). Curbing profanity online: A network-based diffusion analysis of profane speech on Chinese social media. *New Media & Society*, Advance online publication. https://doi.org/10.1177/1461444820905068

Subrahmanyam, K., Smahel, D., & Greenfield, P. (2006). Connecting developmental constructions to the internet: Identity presentation and sexual exploration in online teen chat rooms. *Developmental Psychology*, *42*(3), 395-406. http://dx.doi.org.ezproxy.lib.ryerson.ca/10.1037/0012-1649.42.3.395

Tenove, C., Tworek, H., & Mckelvey, F. (2018, November 12). We can't rely solely on Silicon Valley to tackle online hatred. *The Globe and Mail*. https://www.theglobeandmail.com/amp/opinion/article-we-cant-rely-solely-on-silicon-valley-to-tackle-online-hatred/?__twitter_impression=true

Twitter. (n.d.-a). *Hateful conduct policy*. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

Twitter. (n.d.-b). *You're in control with our safety tools*. https://about.twitter.com/en_us/safety/safety-tools.html

Udupa, S. (2018). Gaali cultures: The politics of abusive exchange on social media. *New Media & Society*, *20*(4), 1506-1522. https://doi.org/10.1177/1461444817698776

Uma, S. (2017). Outlawing cyber crimes against women in India. *Bharati Law Review*, *5*(4), 103-116. http://bharatilawreview.com/uploads/07_Saumya_Uma_103-11611.pdf

Wagner, K. (2017, April 10). Twitter fights to protect anonymous users more often than you'd think. *Recode*. https://www.recode.net/2017/4/10/15244754/twitter-lawsuit-government-anonymous-users

Welsh, A., & Lavoie, J. A. A. (2012). Risky eBusiness: An examination of risk-taking, online disclosiveness, and cyberstalking victimization. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *6*(1), Article 4. https://cyberpsychology.eu/article/view/4260

Wong, J. C. (2018, May 15). Twitter announces global change to algorithm in effort to tackle harassment. *The Guardian*. https://www.theguardian.com/technology/2018/may/15/twitter-ranking-algorithm-change-trolling-harassment-abuse

Author Biographies

Priya Kumar (PhD, SOAS University of London) is a research fellow at the Social Media Lab at Ryerson University's Ted Rogers School of Management in Toronto, Canada. Her current research interests sit at the intersection of data feminism, digital diasporas, and online identity politics.

Anatoliy Gruzd (PhD, University of Illinois at Urbana-Champaign) is a Canada Research Chair in Privacy-Preserving Digital Technologies, an associate professor at the Ted Rogers School of Information Technology Management and the Director of Research at the Social Media Lab at Ryerson University. Gruzd's research broadly explores how social media platforms are changing the ways in which people and organizations communicate, collaborate, disseminate information and misinformation, conduct business and form communities online, and how these changes impact society.

Philip Mai (JD, Syracuse University) is a senior researcher and director of Business and Communications at the Social Media Lab, at Ted Rogers School of Management, Ryerson University, and a co-founder of the International Conference on Social Media & Society. In his work, Mai focuses on tech policy issues, knowledge mobilization, information diffusion, business and research partnerships, and practical application of social media analytics.