Online Harassment

A Form of Censorship

Online Harassment: A Form of Censorship

Published by: SFLC.in

SFLC.in
K9, 2nd Floor, Birbal Road
Jangpura Extension
New Delhi – 14
India

Email: mail@sflc.in
Website: https://www.sflc.in
Twitter: @SFLCin

Table of contents

List of statutes

Constitution of India, 1949

Indian Penal Code, 1860

Code of Criminal Procedure, 1973

Information Technology Act, 2000

Information Technology (Intermediaries Guidelines) Rules, 2011

Universal Declaration of Human Rights, 1948

International Covenant on Civil and Political Rights, 1966

List of abbreviations

APC	Association for Progressive Communications
BJP	Bharatiya Janata Party
CM	Chief Minister
CrPC	Code of Criminal Procedure
DRC	Democratic Republic of Congo
FIR	First Information Report
Govt.	Government
ICCPR	International Covenant on Civil and Political Rights
ICT	Information and Communication Technologies
IDRC	International Development Research Center
IP	Internet Protocol
IPC	Indian Penal Code
IT	Information Technology
LEA	Law Enforcement Agency
NGO	Non-Governmental Organization
NSFW	Not Safe for Work
OHCHR	Office of the High Commissioner of Human Rights
PM	Prime Minister
RSS	Rashtriya Swayamsevak Sangh
SFLC.in	Software Freedom Law Centre, India
SIM	Subscriber Identification Module
SRR	Statement of Rights and Responsibilities
TBTT	Take Back the Tech
TV	Television
UDHR	Universal Declaration of Human Rights
UN	United Nations
UNESCAP	United Nations Economic and Social Commission for the Asia-Pacific
US	United States

Acknowledgements

We express our sincere gratitude to the Software Freedom Law Center, New York, without whose continuing support, this report would not have been made possible. We also thank Ms. Camille Francois (Principal Researcher, Jigsaw), whose inputs provided valuable direction for our research. In addition, we are grateful to all who took time out of their extremely busy schedules to speak with us at length about their experiences and views on this issue. Last but not the least, we thank every stakeholder who attended our roundtable consultations, as they provided us with valuable perspectives that proved critical in furthering our own understanding of the issue.

I. Introduction

The Internet changes with stunning rapidity the modes of communication throughout human society, giving rise to new problems as well as immense opportunities for democratizing educational, commercial and social opportunity. Forms of communication we call "social networking" allow individuals to speak directly to multitudes, and allow multitudes to speak back. This empowers wrong-doers, as well as everyone else. Harms done by aggressive, violent, misogynist and racist speech are increasing in visibility around the world, as the language of the street is amplified and recorded in the Net.

The Internet has been observed to play a prominent role in propagating hateful sentiments during periods of tension as it sets the stage for aggressive campaigning for and against divisive causes, involving elements of intimidation, subdual, fear mongering and provocation among others. Online expressions of political opinions are often met with a flurry of hateful, abusive and harassing responses seemingly aimed at intimidating the speakers into silence, though these have also been known to snowball into acts of real-world violence. For instance, four secular bloggers were murdered in Bangladesh in 2015 by religious extremists, who took strong exception to their writings.

SFLC.in, in our capacity as a non-profit organization seeking to protect and promote civil liberties in the digital world, has been working closely over the years with issues affecting free expression online. Though the Internet has proved over time to be a powerful enabler of free expression, facilitating instant and inexpensive exchange of information across borders, it also presents an equal number of opportunities to those so inclined to transform its speech platforms into vehicles of harassment. Individuals are drawn into digital environments that are increasingly felt as necessary to ordinary social existence, only to find themselves treated as objects in a social theater of aggression and denigration. They often find themselves at the receiving end of sustained abuse, threats and debasement either on grounds of their actual or perceived characteristics, or over their expression of particular ideas and convictions that stand at odds with those of others.

The result, for them, can be as much a form of censorship and silencing as an opportunity for freedom and self-development through learning and expression. The roots of the social phenomenon known here as online harassment lie in another sociological layer from which they have their censoring effects. At SFLC.in, we have been studying online harassment as form of censorship that forces people out of participating in the online discourse. The goal of this study is to explore the phenomenon at both its roots and its surface, to document how people experience its effects in their lives, as well as how the social and technical practices of the platforms establish and differentiate the underlying phenomena leading to abuse. We have been holding public discussions to work with different stakeholders to find workable and understandable ways to address the problem. It must be borne in mind however, that over-reliance on legal-centric responses may come at the cost of collaterally impacting legitimate and permissible free speech, making it essential to emphasize non-state responses to the issue.

1.1 Scope of research

This report will examine the phenomenon of online harassment in a nuanced context i.e. with focus on instances of such speech targeted at individuals over particular ideas and convictions that conflict with those of others. The report will begin with a conceptual exploration of the causes and impacts of harmful speech, and move on to presenting the views of individuals who have been involved in the surrounding debate in some capacity. Next, the report presents the outcomes of two roundtable consultations we organized around harmful speech, and later delve into some state and non-state responses to online hate speech. We will also briefly touching upon a few prominent stakeholders in India that have worked significantly on the issue, and conclude with our observations and recommendations related to online hate speech.

1.2 Methodology

This report has been prepared using a mix of primary and secondary research. A range of existing literature in the form of books, reports, bare texts and commentaries of legislations, articles, and other online resources was consulted and portions have been cited throughout this report. We also spoke with a number of individuals who have been involved in the topical debate in some capacity, so as to present a clearer picture of how mindless abuse and threats of violence affect the usage of the Internet as a speech platform. Views were also solicited from stakeholders during two roundtable consultation conducted on the topic in July and September 2016.

II. Conceptual frameworks

The right to freedom of speech and expression is widely considered one of the most sacrosanct of fundamental human rights, and as such is provided constitutional protection in most jurisdictions. Legislators and jurists have recognized the importance of this right and spoken at length of how it is indispensable to the functioning of democracies. However, freedom of speech also has the distinction of being the right that lends itself most to wanton abuse. The line between legitimate free speech that is merely critical in nature, and potentially dangerous or harmful speech is at times a very fine one, so much so that attempts to prevent the latter can and do spill over to the former.

2.1 The online disinhibition effect

In 2004, Dr. John Suler, professor of psychology at the Rider University, published an article that analyzed a phenomenon known as the "online disinhibition effect" - best described as the nature of Internet interactions in which people say and do things that they never would say or do offline.[1] In his article, Dr. Suler describes two main categories of disinhibited online behavior, namely *benign* and *toxic* disinhibition. The former refers to elevated levels of self-disclosure or kindness online as compared to offline, whereas the latter describes online behavior such as rude language and threats, in which people would not indulge offline.[2] It is pertinent to note however, that this distinction between benign and toxic disinhibition is not always clear. As Dr. Suler points out, a friendly chat might evolve into something more intimate in a way that might cause one of the participants to feel vulnerable or anxious. Similarly, an exchange of seemingly hostile words might be considered perfectly acceptable in certain Internet subcultures.

Furthermore, Dr. Suler outlines six distinct factors as being responsible for the online disinhibition effect. While one or two of these factors cause a predominant share of disinhibition for some, all six factors intersect and interact with each other in most cases, resulting in a more complex and amplified disinhibition effect. The six factors are:[3]

- **Dissociative anonymity:** As the Internet generally allows users to remain anonymous, some users are disincentivized from taking responsibility for their behavior, and instead compartmentalize it in online identities that are kept separate from their offline identities. Anonymity helps users feel less vulnerable about self-disclosing and engaging in antisocial or harmful behavior.

- **Invisibility:** Many, if not most forms of online communications are text-based, and offer a metaphorical shield that keeps users from being physically visible. As a result, inhibitions are lowered because users do not have to worry about tone and body language in their interactions. This lets users misrepresent themselves, as in the case of a man representing himself as a woman, or vice versa, thus presenting possibilities that are not as easily attainable in real world interactions.

- **Asynchronicity:** Some forms of online communications, like email and discussion board

[1] J Suler, *The Online Disinhibition Effect*, Cyber Psychology and Behavior, 2004, Vol. 7, No. 3, p. 321
[2] Ibid.
[3] Ibid., p. 322

interactions, do not happen in real time, which allows users to put up content without immediately seeing any responses from others. This would allow them to gain catharsis and escape any potential negative reactions. As asynchronicity also allows users to think more carefully about what they would like to say before posting, the pressure that can accompany real-life conversations is diminished, enabling them to present differently online than offline.

- **Solipsistic introjection:** In the absence of real-world interactions, online communicators at times assign imagined characteristics shaped by personal expectations and needs to others based on their messages and online persona. Additionally, people sometimes sub-vocalize as they read, which can lead to a perception that they are talking to themselves. This leads to feeling more comfortable talking to the other person and leads to disinhibition.
- **Dissociative imagination:** Online interactions are seen by many as no more than "games" in which the normal rules of social conduct do not apply. This leads some to think that they can adopt and shed certain personas simply by say, logging on and off the Internet. This makes users feel more disinhibited to act in ways that they normally would not offline.
- **Minimization of status and authority:** Authority figures generally express their authority through means such as clothing, body language, name titles, and environments. Even when Internet users are aware of someone's offline status and power, they are less likely to feel intimidated by that authority on the Internet as expressions of the aforementioned cues disappear. The Internet offers a level playing field for all, allowing them to engage with others more as peers instead of as authorities.

While the online disinhibition effect may offer a plausible explanation as to why the Internet plays host to an ostensibly greater number of harmful speech instances as compared to the physical world, it is important to bear in mind that in no way does this detract from the gravity of the problem at hand i.e. harmful speech itself. Leveling abuses and threats, and various forms of debasement, harassment, and incitement to violence among others have all traditionally been treated as punishable offenses across jurisdictions for the reason that speech of this nature is seen as a crime against individuals as well as the collective society, deserving to be penalized, and disincentivized from recurrence. Instances of such speech found online are not made acceptable merely because they occur in the virtual world, and constitute a problem of equal gravity.

2.2 Manifestations of harmful speech online

Harmful speech manifests itself online in a number of ways. The following pages examine the two most common forms of harmful speech found online, namely “hate speech” and “harassment”. Neither form of harmful speech is confined to the virtual world, and both have long been recognized as punishable offenses in the physical world (India's own legal treatment of these offenses is elaborated later on in the report). That said, the contours of online hate speech and online harassment are rather difficult to delineate, specially considering how the latter at times becomes a sub-set of the former.

2.2.1 Hate speech

The term “hate speech” has not yet been defined in a conclusive manner.[4] Due to its complex and interweaving ties with such notions as free expression, individual, group and minority rights, as well as concepts of dignity, liberty and equality, the definition of “hate speech” is often contested. Attempts at arriving at a conclusive definition are made no easier by the recognition accorded under some jurisdictions to the individual’s right to “offend, shock or disturb others”, which would bring numerous instances of what some consider hate speech within the ambit of permissible, legitimate expression.[5] However, also recognized is the fact that some democratic societies restrict expression that spreads, incites, promotes or justifies hatred based on intolerance of actual or perceived character attributes of individuals and communities. Proposed definitions therefore include, but are not limited to speech that advocates, threatens, or encourages violent acts. In common parlance however, definitions of hate speech tend to be broader, extending at times to words that insult those in power, or derogate public figures.[6]

In general, definitions of hate speech take into account some or all of the following components: the content, tone, and nature of speech; the targets of speech; and the potential consequences or implications of the speech act.[7] An oft cited definition of hate speech, as contained in a Recommendation made by the Council of Europe’s Committee of Ministers reads:[8]

[4] Tarlach McGonagle, *The Council of Europe against online hate speech: Conundrums and challenges*, presented at the Council of Europe Conference of Ministers responsible for Media and Information Society (Belgrade, November 2013), p. 4, available at: https://www.coe.int/t/dghl/standardsetting/media/Belgrade2013/McGonagle%20-%20The%20Council%20of%20Europe%20against%20online%20hate%20speech.pdf, last accessed on December 27, 2015

[5] Judgment on the merits delivered by a Chamber, *Handyside v. the United Kingdom*, no. 5493/72, ECHR 1976-II

[6] I Gagliardone, D Gal, T Alves & G Martinez, *Countering Online Hate Speech*, UNESCO, 2015, p. 10, available at: http://www.unesdoc.unesco.org/images/0023/002332/233231e.pdf, last accessed on November 13, 2016

[7] Gavan Titley, Ellie Keen & László Földi, *Starting Points for Combating Hate Speech Online*, Council of Europe, October 2014, pp. 9 – 10, available at: https://www.coe.int/t/dg4/youth/Source/Resources/Publications/2014_Starting_Points_for_Combating_Hate_Speech_Online.pdf, last accessed on December 27, 2015

[8] Council of Europe, *Appendix to Recommendation No. R 97(20) of the Committee of Ministers to Member States on “Hate Speech”*, October 30, 1997, available at:

> *The term ‘hate speech’ shall be understood as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin.*

This is a relatively narrow conception of hate speech in as much as it limits itself to hatred arising from intolerance of certain races, nationalities, and religions. Though this does not necessarily exclude instances where hatred is manifested towards particular individuals on considerations other than racial, ethnic or religious intolerance, the general sense of the term conveyed by its choice of words is that the primary victims of hate speech are larger collectives of individuals such as racial, ethnic or religious groups. An example of hate speech as per this definition would then include, “everyone of [X] faith is barbaric scum and needs to be exterminated”, but not “person [Y] is barbaric scum and needs to be killed”.

A more inclusive definition of hate speech coined by Raphael-Cohen Almagor, reads:[9]

> *Hate speech is defined as bias-motivated, hostile, malicious speech aimed at a person or a group of people because of some of their actual or perceived innate characteristics. It expresses discriminatory, intimidating, disapproving, antagonistic, and/or prejudicial attitudes towards those characteristics, which include gender, race, religion, ethnicity, color, national origin, disability or sexual orientation. Hate speech is intended to injure, dehumanize, harass, intimidate, debase, degrade and victimize the targeted groups, and to foment insensitivity and brutality against them.*

This notion of hate speech is visibly broader, and notably includes within its ambit hostile and malicious speech aimed at individuals due to their actual or perceived characteristics. Going by this definition, the aforementioned statement “person [Y] is barbaric scum and needs to be killed” would fall squarely within the ambit of hate speech, unlike with the Council of Europe's definition. This goes to show that the term “hate speech” comes with a wide range of connotations, depending on who you ask. Some see the term as signifying the expression hateful and disparaging sentiments against racial/ethnic/religious communities with a view to inciting violence against them, while others lend a more liberal interpretation to the term, encompassing all expressions of malicious and intolerant attitudes intended to debase, threaten, and intimidate – be it aimed at communities or

http://www.coe.int/t/dghl/standardsetting/hrpolicy/other_committees/dh-lgbt_docs/CM_Rec(97)20_en.pdf, last accessed on December 27, 2015

[9] Raphael-Cohen Almagor, *Countering Hate on the Internet*, Annual Review of Law and Ethics, Vol. 22 (2014), p. 432, available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2543511, last accessed on June 25, 2016

particular individuals.

Such inconsistencies in understanding the term do make the surrounding dialogue notoriously complex, yet it need not deter the recognition of the problem i.e. the possibility for and the fact of abusive uses of the Internet in its role as a free speech platform.

2.2.2 Harassment

Black's Law Dictionary defines the term "harassment" as "*a course of conduct directed at a specific person that causes substantial emotional distress in such person and serves no legitimate purpose*", or "*words, gestures, and actions which tend to annoy, alarm and abuse another person*".[10] In context of this report, harassment may be understood to signify the persistent use of abusive, threatening, or humiliating content directed at particular individuals either to intimidate and subdue them into silence, or with the intention of discrediting their opinions and views so that they are no longer taken seriously. This may be done by way of persistent emails, chats, text messages, phone calls, and comments/messages on social media platforms, or through more creative means such as setting up fake profiles in their names which are then used to propagate negative content in their names and mislead others into thinking they originated from the targets of harassment.

It is also important to distinguish between the concepts of "hate speech" and "harassment" as these terms are often conflated in relevant policy discussions, leading to much confusion and overlap. While both are widely prevalent forms of harmful speech, they differ substantially in terms of the nature of speech, their respective motivations, and their targets. Simply put, whereas hate speech is generally understood to signify speech that denigrates and/or advocates violence against a racial, ethnic or religious community on the basis of their actual or perceived characteristics, harassment is targeted at individuals rather than groups. Moreover, hate speech is generally considered a more serious offense than harassment, due to its potential to encourage deep-seated prejudices against entire communities, and its tendency to incite large-scale violence against these communities.

2.2.3 Modalities of online harassment

As evident from the preceding paragraphs, the term "online harassment" is not confined in its applicability to harassing words, written or oral, bit is broad enough to encompass a wide range of abusive uses of the Internet as a speech platform. Below are a few practices that have gained notoriety as ways in which individuals can be targeted for online harassment. These may also be considered a typology of key terms to be borne in mind when engaging in dialogues around online harassment.

[10] Working to Halt Online Abuse, Help, available at: http://www.haltabuse.org/help/isit.shtml, last accessed on November 12, 2016

2.2.3.1 Doxing

Doxing (also spelled "doxxing") denotes the practice of harvesting and publishing personally identifiable information i.e. information that can be used either on its own or in combination with other information to identify, locate or contact an individual.[11] This information may be published across a number of online platforms ranging from popular social media platforms to obscure discussion boards, and is sometimes accompanied by express or implied invitations for its viewers to virtually or even physically harass the targets.

While there is no fixed procedure in particular that applies to doxing attacks, social media platforms like Facebook, Twitter, Tumblr, and LinkedIn offer a wealth of personal information such as photos, location and contact details that can be used to easily identify targets and enable others to contact them. It is also possible at times to procure a person's name and home address from a cell-phone number, using such services as reverse phone lookup.[12] Particularly enterprising attackers even use social engineering techniques to extract information from government sources or phone companies,[13] and through domain name and location searching based on the targets' IP addresses.[14]

Once people have been exposed through doxing, they are sometimes shown their details as proof that they have been doxed, in order to intimidate them. Doxing is thus a very popular tactic of online harassment, and was extensively used during the Gamergate and anti-vaccine controversies.[15]

2.2.3.2 Identity theft

Identity theft and identity fraud are terms used to refer to crimes in which someone obtains and wrongfully uses another person's personal data in a way that involves fraud or deception. The ultimate goal of doing so may be to impersonate the victims and leverage their identities to build negative opinions around them, or in other cases to use their credentials for personal profit. Identity theft is a relatively simple affair on social media platforms, requiring minimal effort on part of the perpetrators as all they need is a their targets' names and pictures to create fake profiles. Additionally, information such as national identity numbers, bank account or credit card details,

[11] Adam Clark Estes, *Did LulzSec Trick Police into Arresting the Wrong Guy?*, The Wire, 28 July, 2011, available at: http://www.thewire.com/technology/2011/07/did-lulzsec-trick-police-arresting-wrong-guy/40522/, last accessed on April 20, 2016

[12] Srikanth Ramesh, *What is Doxing and How is it Done?*, 22 March, 2013, available at: http://www.gohacking.com/what-is-doxing-and-how-it-is-done/, last accessed on April 20, 2016

[13] Jason Fagone, *The Serial Swatter*, New York Times, 24 November, 2015, available at: http://www.nytimes.com/2015/11/29/magazine/the-serial-swatter.html?_r=1, last accessed on April 15, 2016

[14] B Blechschmidt, *Guide to doxing: Tracking identities across the web*, 9 November, 2014, available at: https://blog.blechschmidt.saarland/doxing/, last accessed on April 15, 2016

[15] Lance Liebl, *The dangers and ramifications of doxing and swatting*, 28 October, 2014, available at: http://www.gamezone.com/originals/the-dangers-and-ramifications-of-doxxing-and-swatting, last accessed on April 15, 2016

telephone calling card numbers, and other valuable identifying data can also be used, if they fall into the wrong hands, to personally profit at the victims' expense.[16]

The Internet offers several ways for perpetrators to obtain such information. Spyware and other forms of malware may be injected into the targets' devices by inducing them to download files or applications, through email attachments or by having them click on pop-ups, and visit devious websites infected with malicious code. Without the victim's knowledge, spyware runs in the background while it records their Internet browsing habits and keystrokes, monitors the programs they use and collects personal information, which is quietly transmitted to the perpetrator, who can then use the information so gathered to steal money and open credit card and bank accounts, sell it to third parties who will use it for illicit or illegal purposes, and pummel the victim's PC with pop-ups, spam and unwanted messages as well as direct them to websites they never intended to visit.

2.2.3.4 Cyberbullying

Cyberbullying generally refers to the deliberate act of abusing or harassing someone over the Internet. This can be as simple as repetitively sending harassing emails or messages, but may also include such actions as repeated threats, sexual remarks, or defamatory false accusations, vandalizing online content about a person, and posting false statements as fact aimed a discrediting or humiliating the victim. Cyberbullying could be also limited to posting rumors about a person on the Internet with the intention of bringing about hatred in others' minds or convincing others to dislike or participate in online denigration of a target.[17]

Research has demonstrated a number of serious consequences of cyberbullying. For example, victims of cyberbullying have lower self-esteem, increased suicidal ideation, and a variety of emotional responses, such as fear, anger and depression.[18] It has also been reported that Cyberbullying can be more harmful than traditional bullying because there is no escaping it, and that it is an intense form of psychological abuse, whose victims are more than twice as likely to suffer from mental disorders compared to traditional bullying.[19]

2.2.3.5 Cyberstalking

Cyberstalking is the use of the Internet or other electronic means to stalk or harass an individual, a

[16] The United States Department of Just, *Identity Theft*, available at: https://www.justice.gov/criminal-fraud/identity-theft/identity-theft-and-identity-fraud, last accessed on November 15, 2016

[17] Cyberbullying Law & Legl Definition, available at: http://definitions.uslegal.com/c/cyber-bullying/, last accessed on November 15, 2016

[18] Hinduja, S.; Patchin, J. W. (2009). *Bullying beyond the schoolyard: Preventing and responding to cyberbullying*. Thousand Oaks, CA: Corwin Press

[19] *Cyberthreat: How to protect yourself from online bullying*, Ideas and Discoveries: 76. 2011.

group, or an organization. It may include false accusations, defamation, slander and libel. It may also include monitoring, identity theft, threats, vandalism, solicitation for sex, or gathering information that may be used to threaten or harass. Cyberstalking differs from cyberbullying in that the former is often accompanied by real-time or offline stalking.[20] However, some argue that cyberstalking is a form of cyberbullying, and the terms are often used interchangeably.

Cyberstalking has increased exponentially with the growth of new technology and new ways to stalk victims. In January 2009, the Bureau of Justice Statistics in the United States released the study "Stalking Victimization in the United States," which was sponsored by the Office on Violence Against Women. The report, based on supplemental data from the National Crime Victimization Survey, showed that one in four stalking victims had been cyberstalked as well, with the perpetrators using Internet-based services such as email, instant messaging, GPS, or spyware. The final report stated that approximately 1.2 million victims had stalkers who used technology to find them.[21] The Rape, Abuse and Incest National Network, in Washington D.C. has released statistics that there are 3.4 million stalking victims each year in the United States. Of those, one in four reported experiencing cyberstalking.[22]

2.3 Actual incitement to violence

It has long been postulated that there exists a causal link between online harmful speech and real-world violence. For instance, a 2014 report by Muslim Advocates, a San Francisco based legal and advocacy group, found that anti-Muslim hate speech is commonplace on the Internet and can motivate some people to commit acts of violence against Muslims.[23] The report said that anti-Muslim websites give like-minded people a place to gather and at the same time win new supporters through their posts. A cited example was the Facebook page of anti-Muslim blogger Pamela Geller, which grew from roughly 19,000 followers in July 2013 to 78,000 in April 2014.[24]

Despite the relative absence of concrete evidence supporting this causal link, harmful speech, especially when manifested as hate speech, is a key indicator of probable mass atrocities. However,

[20] Brian Spitzberg, Gregory Hoobler, *Cyberstalking and the technologies of interpersonal terrorism*, New Media & Society, February 2002, pp. 71–92, available at: http://www-rohan.sdsu.edu/~bsavatar/articles/Cyberstalking-NM&S02.pdf, last accessed on April 20, 2016

[21] Christa Miller, *High Tech Stalking*, 1 May, 2009, available at: http://www.officer.com/article/10233633/high-tech-stalking, last accessed on April 20, 2016

[22] Tom Smith, *Criminals use technology to track victims*, 28 February 2010, available at: http://www.timesdaily.com/archives/article_4ff0ea3e-4f84-5888-bdc7-b41c03ac9434.html, last accessed on April 20, 2016

[23] Omar Sacirbey, *REPORT: Internet hate speech can lead to acts of violence*, The Washington Post, 6th May 2014, available at: https://www.washingtonpost.com/national/religion/report-internet-hate-speech-can-lead-to-acts-of-violence/2014/05/06/9a9d9e60-d52c-11e3-8f7d-7786660fff7c_story.html; the complete report is available at: http://www.muslimadvocates.org/wp-content/uploads/Click-Here-to-End-Hate.pdf, last accessed on April 20, 2016

[24] Ibid.

Professor Susan Benesch of the Berkman Klein Center for Internet and Society, Harvard Law School, suggests that hate speech i.e. speech that denigrates people on the basis of their membership in a group, such as an ethnic or religious group, is too broad for successful early warning of mass atrocities for two related reasons:[25]

- Hate speech is common in many societies, including those at minimal risk of violence.
- Some hate speech does not appreciably increase the risk of mass violence, although it may cause serious emotional and psychological damage.

For these reasons, Prof. Benesch proposes the use of the term "dangerous speech", to be understood as a speech act that presents a reasonable chance of catalyzing or amplifying violence by one group against another, given the circumstances in which it is made or disseminated.[26] She further proposes a set of guidelines comprising five variables, which are expected to help identify instances of dangerous speech in time to serve as early warnings of mass violence. The most dangerous speech act, or ideal type of dangerous speech, would be one for which all the following five variables are maximized:[27]

- *The speaker*
    - Did the speaker have authority, power or influence over the audience?
    - Was the speaker charismatic or popular?
- *The audience*
    - Who was the audience most likely to react to the speech at issue?
    - Was the speech directed primarily at members of the group it purported to describe, i.e. victims, or at members of the speaker's own group, or both?
    - Did the audience have the means or capacity to commit violence against the group targeted in the speech?
    - Was the audience suffering economic insecurity, e.g. lacking in food, shelter, employment, especially in comparison with its recent past?
    - Is the audience characterized by excessive respect for authority?

[25] Susan Benesch, *Dangerous Speech: A Proposal to Prevent Group Violence* (2013), p. 1, available at: http://dangerousspeech.org/guidelines, last accessed on January 6, 2016

[26] Ibid.

[27] Ibid.

    - Was the audience fearful?

- *The speech act*
    - Was the speech understood by the audience as a call to violence?
    - Did the speech describe the victims-to-be as other than human, e.g. as vermin, pests, insects or animals?
    - Did the speech assert that the audience faced serious danger from the victim group?
    - Did the speech contain phrases, words, or coded language that has taken on a special loaded meaning, in the understanding of the speaker and audience?
- *Socio-historical context*
    - Were there underlying or previous conflicts between relevant groups?
    - Were there recent outbreaks of violence following other examples of hate speech?
    - Were other risk factors for mass violence present, such as weak democratic structures and rule of law, and structural inequalities and discrimination against a group or groups?
- *Means of transmission*
    - Was the speech transmitted in a way that would reinforce its capacity to persuade, e.g. via a media outlet with particular influence or without competitors?
    - Was the audience exposed to, or did it have access to, alternate views or sources of information?
    - Was the speech frequently repeated, in similar form or content?

The dangerous speech framework serves an as excellent frame of reference for evaluating particular speech instances against their likelihood of provoking large-scale violence. However, as the framework was designed to apply specifically to speech directed at a collective audience rather than individual recipients, its applicability to the object of this report i.e. online harassment is somewhat limited. Variables such as the authority and charisma of the speaker, the audience's capacity for violence, the nature of the speech act and the availability of alternate sources of information, are nevertheless relevant factors to consider when evaluating harassing speech for its potential to trigger violence against its targets.

III. Stakeholder perspectives

In order to get a clearer understanding of the real-world consequences of online harassment, SFLC.in spoke to 17 individuals, who had been involved in the topical debate in some capacity. Not entirely by coincidence, a number of interviewees have been associated, in the past or currently, with the field of journalism and have been quite forthcoming with their independent and bold views on the most relevant topics of the day. As a result, these individuals face frequent threats, abuses and other miscellaneous verbal disparagements in the course of their online activities, making them ideal candidates to speak from the perspective of targets of online harassment. The list of interviewees also features legislators, civil society actors, and other expert stakeholders, who by virtue of their frequent reliance on the Internet, have had at least first-hand glimpses at the plight of the harassed.

During the interviews, questions were asked on specific instances of abuse; who were the people/accounts involved in the intimidation campaign and what were the emotional and psychological after effects of the incidents. To understand the long and short term effects of harassment, questions were posed on whether they felt safe and free to express their opinions online anymore, and whether they were happy with the action taken by the platform on their complaints and the features that were provided by it to deal with harassment. Interviewees were asked about whether they ever considered taking legal action and if so, how successful and easy was the process. In most cases, the complaints never proceeded beyond the preliminary investigations, after which it was dropped. To understand what could be done to better the situation, opinions were collected from the interviewees on what kind of measures could be taken by the platform and Law Enforcement Agencies (LEAs) to better the situation and also what is the desired cultural and attitudinal change in the public. We wish to impress upon the reader at this point that the following conversations do not make for an adequately representative data pool that enables drawing broad-based conclusions about the nature of the problem or its typical consequences. These should be seen rather as first-hand accounts from individuals who have personally experienced targeted online hate campaigns, thereby serving to humanize such targets who are often seen as distant online entities by their attackers.

3.1 Abhinandan Sekhri; Co-Founder, Newslaundry

Abhinandan Sekhri is a journalist and co-founder of Newslaundry, an online media venture that focuses on news critique, news reports, and current affairs. As a journalist and an entrepreneur with a web venture, he admits to spending every day except the weekends on Twitter and Facebook, trying to get information, as well as publicizing his stories and views. In his opinion, online forms of media trumps the traditional mediums of communication, such as newspapers and televisions by not only eliminating the requirement of large capital and investment, but by also enhancing the scope of engagement by being a two way channel, where both the producers of content as well as the consumers can interact and engage. Nevertheless, with respect netizens' behavior, Mr. Sekhri believes that the online is simply an extension of the real physical world.

Being a journalist who expresses his views quite often on social media platforms, Mr. Sekhri has received numerous messages and comments that could be termed as trolling, abuse, and harassment. Although he condemns this behavior, he doesn't think it is worthy of punishment. However, while admitting to the long rope he provides to freedom of expression, he deliberately makes a distinction between the kind and extent of abusive messages that he receives vis- a- vis the “viciousness” of such comments received by women on these platforms. Though Mr. Sekhri does accept that anonymity is a feature that shields people while they say the vilest and the most disgusting things, he does not attribute the increase in such behavior on online platforms solely to this one characteristic. He maintains that there are other aspects of online media, such as the lag time in reaction, the inability to see the user in person, or hear one's own voice when typing, that makes it easier to make abusive and/or harassing statements on online platforms.

Citing a personal experience to debunk the emphasis put on anonymity being the vehicle of harassing speech, he recalled that a childhood friend had made an abusive remark on a Facebook post of his, that Mr. Sekhri was sure he would have never made in person. On the aspect of being a target for abuse, he narrated an incident where someone said, “Azaadi Azaadi, you support these Azaadi valaas, kissi din beech chowk main khade ho ke koi tumhe azaadi de dega” (*Freedom freedom, you support these advocates of freedom, some day we will give you “freedom” in the middle of the street*) and then he put an icon of a gun and a bomb, and Mr. Sekhri replied that, “Tumhi aa ke Azaadi de do, agar tum main itni himmat hai, yeh anonymous twitter handle se kya dhamki de rahe ho” (*If you have the courage, why don't you yourself come and set me 'free', why are you hiding behind an anonymous Twitter account*).

Even though he maintains that his usage or behavior is not affected by such comments or attacks, there are friends he knows who have left the platforms because of extensive abuse. Moreover, he strongly holds the opinion that men and women are subjected to different kind of harassment in the country due to traditional roles and socio-cultural history. Where women in India are subjected to more sexually demeaning and illicit comments owing to the country's societal structure, similarly, racial hatred is more prominent in terms of harassment in the West.

Talking about reporting mechanisms employed by the social media platforms as well his usage of these tools, Mr. Sekhri exclaimed that he has never reported anything to either Facebook or Twitter (two of his most used platforms in online interactions), nor does he know anyone personally who has resorted to these means. He justified the non usage of these tools by stating that amidst the vast expanse of abusers, mob attacks, and troll armies, how many of those Twitter handles (or Facebook accounts) would one keep reporting to these intermediary platforms; with quite a few of these being anonymous handles.

While explaining shortcomings of the present system established on these platforms to tackle these abuses, he opined that it is humanly impossible to sift through the enormous volumes of content generated on these platforms on a daily basis. An algorithm would not suffice owing to the complexity of the messages that need to be deciphered to reach a conclusive decision on their hateful character; therefore, human intervention in the process is necessary, and swift action on all content is hence slightly far fetched. He also pointed out the paradox of anonymous speech, and exclaimed his helplessness to find a solution where it could be utilized for exercising freedom of expression, but is not exploited for issuing violent threats and meting out harassing and abusive messages.

The liability of platforms, according to Mr. Sekhri is akin to managing a virtual public space. He compared this with an entertainment park and remarked that although the owners can't throw anyone out because they don't like them, they do need to be able to control the environment they have created. Moreover, this discretionary power (of intermediaries) has to be scrutinized carefully for reasonableness of actions as well. On being asked about approaching LEAs, Mr. Sekhri questioned the utility of that action and dismissed it as being futile as according to him, there is an extremely small percentage of the economically empowered community that approaches law enforcement officials, and it is only the people with little to no means and who do not have any 'connections', that move to the law enforcement for their help. This attitude entails a lack of faith in their services and capability, and therefore, lesser reliance is placed by the community in general on LEAs in such situations.

Speculating on the reactions of the police force to complaints filed for harassing or abusive behavior online, he exclaimed that their reaction will be determined based on the the status of the person making the complaint. In his opinion, a common man making a complaint against abuses being said to him on social media will not lead to any fruition.

As suggestions for limiting the abusive or harassing speech online through reforms on the regulatory, legal, or user level, Mr. Sekhri remarked that for extremely violent posts and messages, public shaming might help, but this solution is not viable for anonymous users. However, he believes that the problem will solve itself in due course of time, as the novelty of the medium fades away, people will understand that being abusive can only gain them attention to some extent, a more sensible use of social media will gradually emerge and sustain. Moreover, Mr. Sekhri disapproves of a chilling effect such abuse might cause on the first time users of the Internet because in the long run, according to him they will realize that the online world is simply an extension of the offline world and the benefits of the Internet overpower the anger and hate that is generated by a few on these platforms.

3.2 Arvind Gupta; National Head, Information and Technology, BJP

Dr. Arvind Gupta is the National Head of Information and Technology at BJP, and is also the founder of the Digital India Foundation. He spearheaded the Digital and Social Media campaign for Prime Minister Narendra Modi during the 2014 Elections, and is also well-known as an Innovation Evangelist with over 20 years of experience in leadership, policy, and entrepreneurial roles. Please note that the views expressed here are strictly his personal views, and must not be seen to represent those of the BJP itself.

According to Dr. Gupta, he and the BJP believe in a "DigitalFirst" philosophy when it comes to online media. The idea is not only to bring parity in information dissemination, but also to look at overall parity so that everybody gets information together. This makes information available to all in real time and in the right manner, enabling a truly digital democracy. In this regard, Dr. Gupta feels that the network effect and engagement of social media cannot be compared to electronic or print media.

Dr. Gupta identifies at least three factors that motivate people to behave differently online than they do offline. The first of these factors is anonymity i.e. the option of conducting oneself online under a fictional identity or without an identity altogether; the second, accessibility i.e. the availability of affordable, widespread access to the Internet and Internet-related services for a highly eclectic user-base; and the third, reach i.e. the possibility for communicating with a vast and immediately unquantifiable audience. In other words, as the Internet allows its users to anonymously and cost-effectively interact with a large number of similarly placed users, human interactions work somewhat differently online than offline, where one or more of these factors are generally absent.

Being a public figure, to whom extensive use of the Internet is a core component of his work, Dr.

Gupta reports having experienced these variations in human interactions first-hand on numerous occasions. He has been targeted online for his views many times in the past, often in a harassing manner. He has also been trolled and threatened many times during his use of the Internet as a speech platform. Though Dr. Gupta feels that it is sometimes better to switch-off and simply not respond to instances of online harassment, he also admits that this is much easier said than done. A relentless stream of negative content thrown one's way can and will disrupt one's online activities and impact them in appreciably harmful ways. While he does agree that online harassment can have a silencing effect in the short term, he remains optimistic that this is not a long term effect. He also notes that the silencing effect is much more perceptible in public online presences than in anonymous users.

As for the content reporting and removal mechanisms currently available with popular online speech platforms, Mr. Gupta's experience using them has been mostly disappointing. He finds that the platforms themselves have very poor response mechanisms. Though he has on several occasions personally reported harassing content that clearly violated applicable content policies, in his experience, the action taken by the platforms is often too little and too late. There have even been instances, where such content reports have not elicited any response from the platforms at all. Dr. Gupta feels that platforms need to urgently address these issues to remain relevant. Moreover, this must be done very independently and without political or affiliation partisanship. Online platforms presently suffer from lack of trust when it comes to guaranteeing users' safety, and the significant levels of human intervention in content moderation further dilutes this trust as it involves personal biases. As a response, Dr. Gupta is of the opinion that platforms need to build a system of trust and non partisanship by heeding user feedback and implementing broad-based changes to their content moderation practices on the basis of this feedback.

Dr. Gupta also finds that law enforcement officers are often not adequately trained to register and investigate cyber grievances, and usually function in a state of under-preparedness. Police officers are frequently observed to be unaware of efficient models of redressing such grievances, which when combined with the poor response mechanisms available with online platforms, leaves users with little recourse when faced with online harassment. Dr. Gupta recommends focused capacity building exercises involving regular training for law enforcement officers as a great starting point for rectifying this state of affairs.

3.3 Baijayant Panda; Member of Parliament, Lok Sabha

Baijayant 'Jay' Panda is a serving Member of Parliament (Lok Sabha) widely known for his campaigns supporting key public causes such as public health and migrant labor. He has been a vocal proponent of respect for free speech at all levels of governance, and has tabled several Private Member's Bills before the Indian Parliament to remedy laws that restrict it. Mr. Panda uses the Internet on a daily basis to express his views on relevant issues of public import and to interact directly with citizens, and believes the medium offers two primary advantages over traditional media like print and television. First, the Internet has significantly greater reach than its traditional counterparts, in that it allows users to instantly and inexpensively communicate with a global audience. Second, it generally allows recipients of communications to send feedback directly to the communicators, thus making it a truly interactive medium.

Being a prominent political figure who is also a frequent user of social media platforms like Twitter and Facebook, Mr. Panda routinely receives negative remarks in response to his own public communications. These remarks range in severity from frivolous and factually inaccurate aspersions to hateful and threatening messages, and their sources are not always tied to any particular ideologies or political outlook. While his usual response to such content is to simply ignore them, he does mute content that is clearly identifiable as spam i.e. when the originators are identifiable as not real people. It is very rarely that Mr. Panda blocks people on social media, but he is at times forced to go to the extent when faced with particularly vile abuse, though this is certainly never done over mere disagreements. Reporting users to the platforms too is an option rarely invoked by Mr. Panda, except with repeated and vexatious trolls. He refrains from using content reporting tools like the ones mentioned above as he finds them more or less ineffective in the long run. On Twitter for instance, he pointed out that blocking a user does not hide retweets of the blocked user's content.

For this reason, Mr. Panda believes there is a need for adequate legal protections against online abuse, including online harassment. He emphasized that this should not be seen under any circumstance as an endorsement of draconian laws like the now-repealed Section 66A of the Information Technology Act, 2000 (IT Act), which lent itself to wanton abuse due to its over-broad and ambiguous language. The need of the hour rather, is a framework of complementing laws that reflect consistency in their outlook across statutes. This framework must respect freedom of speech at all levels, as this is a Fundamental Right that is indispensable in democracies. Mr. Panda also cautions against the retention of outdated laws that fail to evolve with changing times. Section 124A of the Indian Penal Code, which prescribes severe penalties against the offense of sedition was highlighted as an example of an outdated law, as it punishes mere ideas and views expressed against the State, rather than actions or incitement to imminent violence.

Mr. Panda stops short of recommending self-regulation by online speech platforms as the sole defense against harmful and harassing speech. While he admits that most platforms rightly have low levels of tolerance towards such content, he finds that the collective wisdom of the nation embodied in its laws will always be a better judge of permissibility. For this reason, content restrictions imposed by online platforms through content policies may be seen as the first line of defense against online harassment, but they can never be the sole line of defense.

Mr. Panda also recognizes that there exist certain shortcomings with the existing state machinery when it comes to tackling online harassment, including under-prepared law enforcement officials. However, their state of under-preparedness is also understandable considering they are burdened with the daunting task of ensuring citizens' safety offline as well as online. To overcome these shortcomings, Mr. Panda recommends greater emphasis on capacity building exercises. Some such exercises are already underway, but this must be an ongoing process that follows a consistent and regimented approach. There may even be a case to be made for specialized cells that deal exclusively with matters relating to online hate and harassment, as the scale of the problem certainly warrants it. Aside from the above, Mr. Panda also recommends focused initiatives aimed at generating multi-stakeholder awareness around the issue, including greater levels of cooperation between the Government and civil society, and periodic publication of information materials such as explainer videos and do's and dont's for online conduct.

3.4 Bishakha Datta; Executive Director, Point of View

Bishakha Datta is a film maker, activist and a former journalist. She is the co-founder and Executive Director of Point of View, a non-profit organization based in Mumbai, working in the area of gender, sexuality and women's rights. Point of View was involved in organizing a workshop in 2013 aimed at discussing means to resist undue content regulation and generating awareness about online security and privacy. The organization has also worked on and published articles and reports on gender abuse online.

While Ms. Datta has not been at the receiving end of any targeted hate campaigns unlike most other interviewees on our list, she does not hesitate to agree that there indeed are numerous lynch mobs on social media platforms like Twitter. She views Twitter as a platform for political rather than personal expression, and believes there has been an overabundance of reports of people being attacked online, more often than not by supporters of right-wing politics. Though this prevalent atmosphere of intolerance and abuse has not had a silencing effect on her personally, Ms. Datta believes it would be unreasonable to expect everybody to toughen up and cope well with mindless abuse.

The need of the hour, says Ms. Datta, is to institute easier and more transparent processes to complain to the online platforms. In cases of online hate, approaching law enforcement should ideally be a last resort for when there are direct threats to one's safety. There are many support groups and other such organizations that fight online hate and abuse at the individual level, but the social media platforms must also recognize that they also enable those that tend to take away others' right to free speech. By extension, they need to offer easy mechanisms of reporting by which hateful and abusive behavior can be identified and dealt with. If these platforms are unable to create spaces for its users while at the same time ensuring the users' safety, why would anyone use their services to begin with? Of course we need laws, but we also need easy measures that people can exercise at the platform level.

Ms. Datta feels that people still think of the physical world as the real world and they don't consider digital world to be very “real”. This perception puzzles her because people spending so much time online has obviously eroded the boundary between online and offline to a great extent. People are forming intimate relationships and friendships online; they announce important personal milestones online; to pretend that none of this happens in the “real world” does not make sense. Moreover, online actions can have very real offline consequences - we have had many cases of young women being bullied online; there have been cases both inside India and outside of people committing suicides because of such bullying; people have gone into deep mental depression. We cannot say that it's any less harmful than physical harm, or any less real to that person.

The Internet may thus have offered new avenues for the expression of hate and intolerance, but this is not to say that it outweighs the benefits offered by the Internet. According to Ms. Datta, one crucial difference between online and offline speech is that the former accommodates a great diversity of voices. While mainstream media will always present content from a colored perspective, online media permits everyday citizens to express their legitimate views without censorship, thus enabling them to influence public perception on issues in their own capacities. For instance, in connection with her work, Ms. Datta follows a number of sex-workers on Twitter, who would normally have been isolated from the general public. She finds it very interesting to see how they have a voice on the Internet, which they just don't have offline. They admittedly must face considerable abuse online, yet they are also able to stand up for themselves present their views, which makes the Internet a great torchbearer of modern democracy.

3.5 Hartosh Singh Bal; Political Editor, The Caravan

Hartosh Singh Bal presently works as the Political Editor of the Caravan – a national politics and culture magazine, and has been embroiled in his fair share of public controversy over the years. Mr. Bal has been openly critical of several right and left wing politicians and political parties, notably of Prime Minister Modi (then Chief Minister of Gujarat) for his handling of the 2002 Gujarat riots and of the Indian National Congress for their handling of the 1984 anti-Sikh riots. In November 2013, he was removed from his position as Political Editor of the Open Magazine – a general interest and current-affairs publication, as he was considered to have made several political enemies on account of the views expressed in his writings and in his television appearances. As an active Twitter user with more than a few unpopular opinions, Mr. Bal regularly receives abusive and threatening messages over the platform.

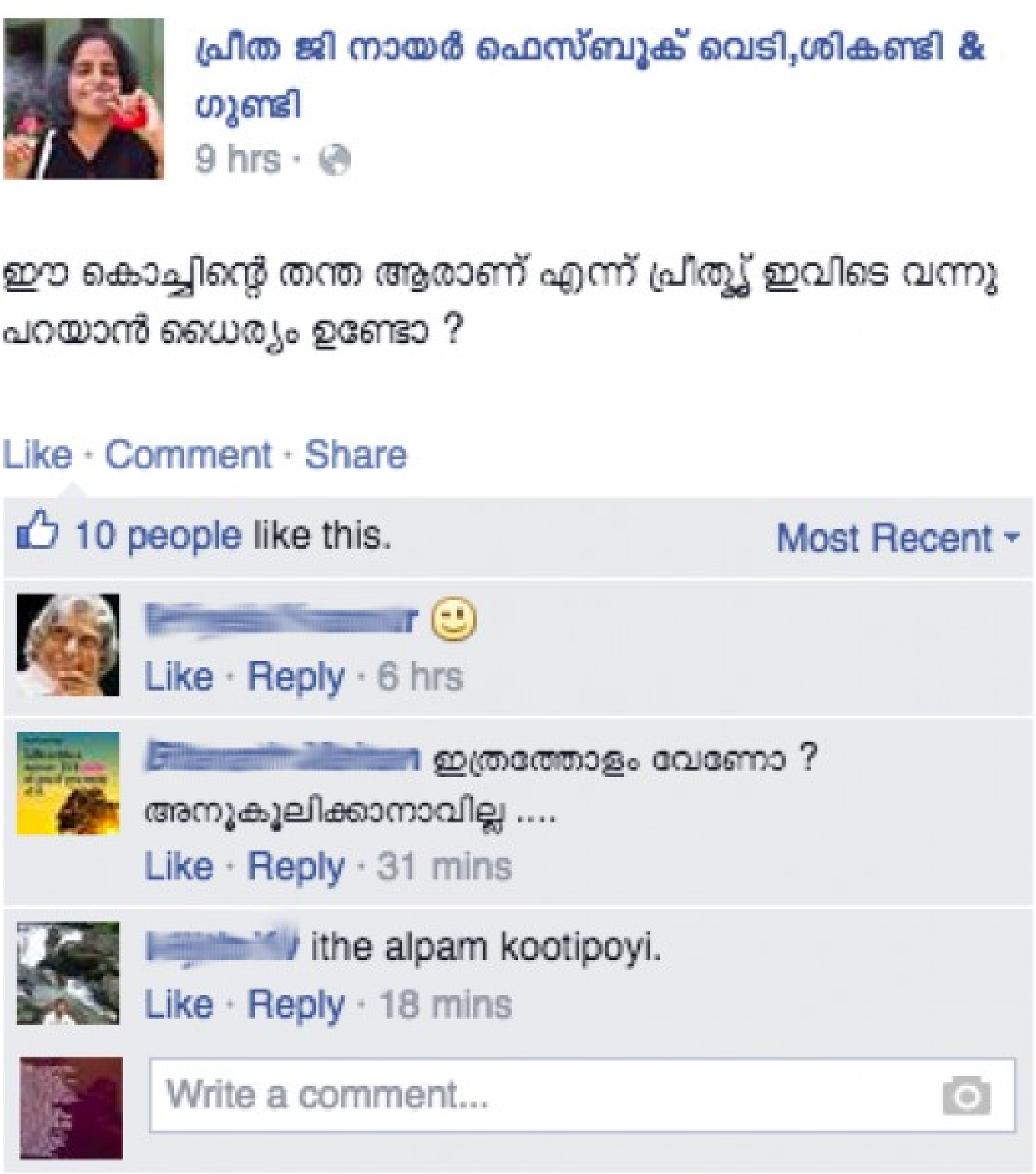
Speaking to SFLC.in about engaging with his online abusers, he said that in his experience, most of the abusers eventually back off. Even when they don’t, he finds most attacks to be senseless tirades that can be handled simply by having a thick skin. While he has faced all kinds of abuse online, including name-calling, disparaging remarks against his family and ancestors and so on, he finds the charges against him so idiotic that none of the abuses really bother him or have a silencing effect on him. He notes that this might in part be due to the nature of his social media use, in that he uses various platforms to voice his own opinions rather than as a means to stay updated on what is going on elsewhere.

While Mr. Bal finds blocking, muting, and other such reporting mechanisms to be easy options available to the victims of online abuse, he rarely exercises these options himself. As far as he is concerned, the stupider his abusers look on social media, the better. For him, the rampant abuse only underscores how unorganized and mindless India’s right wing social media trolls truly are. As regards the possible existence of an “abuse syndicate” (referring to organized and possibly politically funded online groups whose mission is to debase, threaten and harass those expressing conflicting political views so as to silence these voices) as speculated by other interviewees, he finds the notion plausible since many such accounts are dubious and feature sporadic coordination and activity at best. However, he refrains from making any conclusions as there is no real evidence that points to the existence of such a syndicate.

Despite his countless experiences with hate and intolerance online, Mr. Bal believes that with the nature of social media, the latitude for speech should be very broad. Its only when speech clearly crosses the line, when the harassment is personal, directed at faith or is physically threatening, that the platforms themselves should step in. In his words, he would rather err on the side of latitude than on the side of caution despite all the problems.

3.6 Inji Pennu; writer, activist

Writer Inji Pennu spearheaded a popular campaign against Facebook's infamous "real name policy" in late 2015, when the Facebook accounts belonging to several women (including herself) were suspended as result of malicious and vengeful complaints regarding their use of inauthentic names on the social networking platform. The motivation behind her campaign were the sustained hateful attacks on Facebook against one Preetha G Nair – an activist from Southern India who also happened to be a single mother – due to her linking an article that spoke of the recently deceased former President of India A P J Abdul Kalam's sympathies for right-wing ideologies. Soon after Preetha posted the link, Facebook erupted with relentless and vile abuse, some calling her a slut and an abomination, others going so far as to question the parentage of her autistic son.



പ്രീത ജി നായർ ഫെസ്ബുക് വെടി,ശികണ്ടി & ഗുണ്ടി
9 hrs ·

ഈ കൊച്ചിന്റെ തന്ത ആരാണ് എന്ന് പ്രീതയ്ക്ക് ഇവിടെ വന്നു പറയാൻ ധൈര്യം ഉണ്ടോ ?

Like · Comment · Share

10 people like this. Most Recent

[redacted] 🙂
Like · Reply · 6 hrs

[redacted] ഇത്രത്തോളം വേണോ ? അനുകൂലിക്കാനാവില്ല ....
Like · Reply · 31 mins

[redacted] ithe alpam kootipoyi.
Like · Reply · 18 mins

[*Translated: Can Preetha tell us here who the father of the child is? (Comments say the question crossed a line)*]

Ms. Pennu, when alerted of the incident through her Facebook network, initially waited for the attacks to die down as she presumed they always do. However, as the hate campaign against Preetha showed no signs of slowing even after a week, she authored and published a blog-post about it through Global Voices – a non-profit organization that works on issues affecting the marginalized by means of citizen reportage. As soon as her blog-post went public, the hate campaign turned against her. The same people who had been attacking Preetha began attacking her, though the attacks were relatively muted as she had always been cautious of the personal information she shared publicly.

Things soon took an interesting turn as the spiteful attackers reported the accounts of Preetha, Ms. Pennu, and a number of women who had publicly protested the attacks to Facebook as using "fake" names on their profiles. As a result of these reports, Facebook in exercise of its real name policy, promptly suspended the user accounts of all involved until such time as they furnished documents verifying their identities. Ms.Pennu was understandably outraged by what she felt was an extremely arbitrary and unjust policy that endangered all those suppressed and marginalized individuals who used alternative names on Facebook for legitimate safety concerns. She enlisted the help of supporters including Global Voices and the Electronic Frontier Foundation to launch a campaign in protest, during which she participated in numerous meetings with Facebook representatives, who initially were unwilling to do anything about the policy, claiming it to be an essential safety feature that introduced an element of accountability to the system.

As Ms. Pennu's campaign gained momentum and more individuals and organizations across the world signed on in support, Facebook partially relented to the mounting pressure and undertook to make amendments to their policies and elaborate the real name requirement in more descriptive and liberal terms. Changes made to Facebook's Community Standards in December 2015 included language which clarified that individuals were not under an obligation to supply their legal names, rather the names they associated themselves with in every-day life, even if they were not the same as the names on their identification records. The real name policy however continues to be applicable in spirit, and user accounts reported as using fake names continue to be suspended without notice until their authenticity is definitively proven. Ms. Pennu vows to keep fighting the policy until it is abandoned entirely, and her Facebook account remains suspended to this day as she refuses to provide her identification records because she feels Facebook's security protocols are not strong enough to guarantee their complete safety.

Looking back on what started it all, Ms. Pennu feels that the general society in Kerala, the Southern Indian state from where she and Preetha hail, boasts multiple layers of ingrained misogyny, which is what motivated the attacks against Preetha and her to begin with. She thinks that despite being known for being the most literate state in India, and perhaps because of it, Kerala’s men folk intelligently turn everything into their own misogynistic lines of thought.

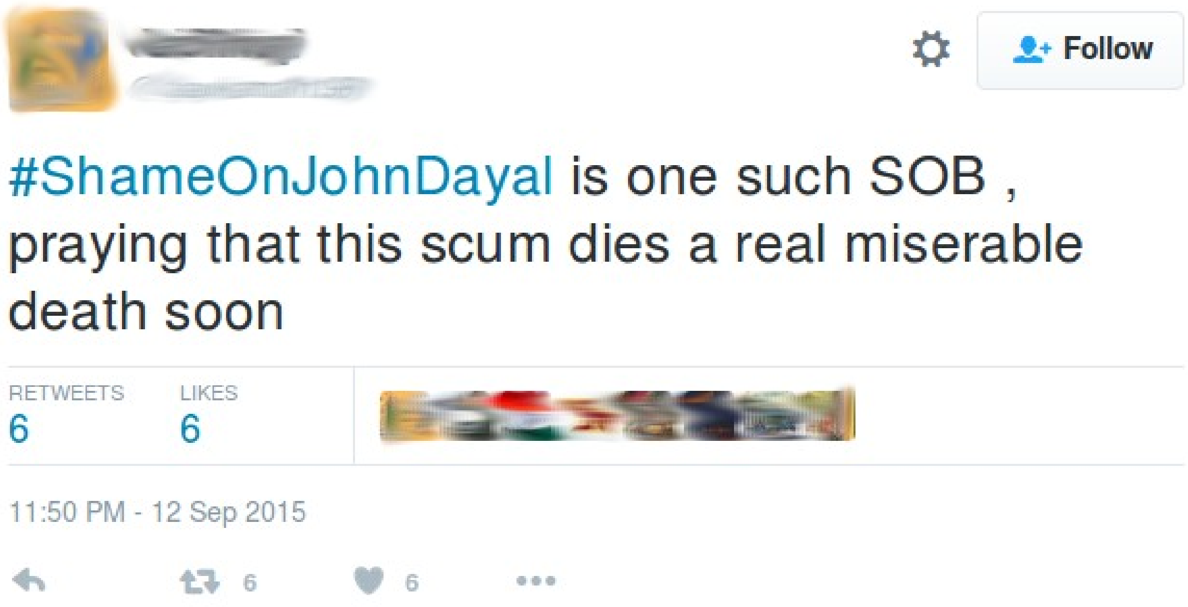
She lamented the fact that the most common response received from the police when they are approached with complaints regarding attacks such as the ones in her and Preetha’s case is to simply not use Facebook or any social media. She pointed to the experience of one of her friends, who was told by the police to button up her shirt before making complaints, as illustrative of the levels to which misogyny has permeated in Kerala’s social setup. According to Ms. Pennu, the police also fails to see how online threats can cause serious, even fatal damage, but chooses instead to believe that they disappear as soon as the computers are powered down.

Ms. Pennu believes with good reason that our future will be an entirely digital one, where the lines that separate our online and offline lives will be blurred into non-existence. She feels that policymakers must therefore start giving more importance to the digital world and think about how a mob could attack somebody in the virtual world, just as effectively as in the physical world.

Any campaign, she remarks, is all about creating awareness, so that when products are designed, they pay due attention to all the relevant issues. If nothing else, Ms. Pennu is comforted by the fact that people are beginning to speak up and stand fearlessly for what they believe is right.

3.7 Karuna John; freelance journalist

In September 2015, John Dayal, a well-known minority rights activist (and Karuna John’s father), observed on Twitter that a local school, where a student had recently been raped, was owned and operated by members of the Bharatiya Janata Party (BJP). He pointed out that the incident had not received much attention from the media, and speculated that BJP’s role in the institution’s administration might have had something to do with this. Foreseeably enough, Mr. Dayal’s observations did not sit well with numerous BJP loyalists, and a stream of abuses and violent threats followed. A particularly enterprising abuser managed to get hold of Mr. Dayal’s contact details, promptly doxed him, posting the information on a public forum and inviting others to threaten and abuse Mr. Dayal.



[redacted]

#ShameOnJohnDayal is one such SOB , praying that this scum dies a real miserable death soon

RETWEETS 6 LIKES 6

11:50 PM - 12 Sep 2015

[*One of the many harassing tweets with the trending #ShameOnJohnDayal tag*]

Recalling the incident, his daughter, Ms. Karuna John said, “Someone put his personal phone number out there and that is when things got messy. Usually we would ignore something like this, but father was retired at the time and quite old. Neither me nor my brother lived with our parents. So I quickly alerted my immediate family and got him to switch off his phone. But by that time he had already got a couple of abusive calls from random people. The issue had also started trending on Twitter with so many abuses flowing in, some asking to eliminate him.”

Ms. John informed us that due to the nature of their work and the unpopular opinions they frequently espoused, both she and her father were no strangers to abusive/threatening messages over social media, with she frequently receiving sexually explicit abuses. However, this was the first time either of them had faced a doxing attack, so they consulted some friends and lawyers on what needed to be done. As the incident was reported by some media outlets, several individuals expressed solidarity with the family, and a petition was started to put an end to the continuing abuse.

On approaching the police and filing a First Information Report (FIR) the next morning, Mr. Dayal was informed that it would be difficult to trace the perpetrators, though Ms. John remained skeptical of their inability seeing how the same officers had no trouble taking notice of “objectionable” Facebook posts and making arrests - as they did for instance, when two girls from Maharashtra were charged under Section 66A of the IT Act for questioning the city's shutdown over a local politician's death. Ms. John was equally disillusioned by Twitter’s own redressal mechanisms as no action was reportedly taken despite her having filed numerous complaints. Till date, she has muted and blocked almost a thousand abusers on Twitter, yet new messages pour in everyday. Sometimes the abuses are not even concerned with what she has personally said or written, and instead target her based on her association by way of work, religion, or gender with other individuals who might have said/done something objectionable.

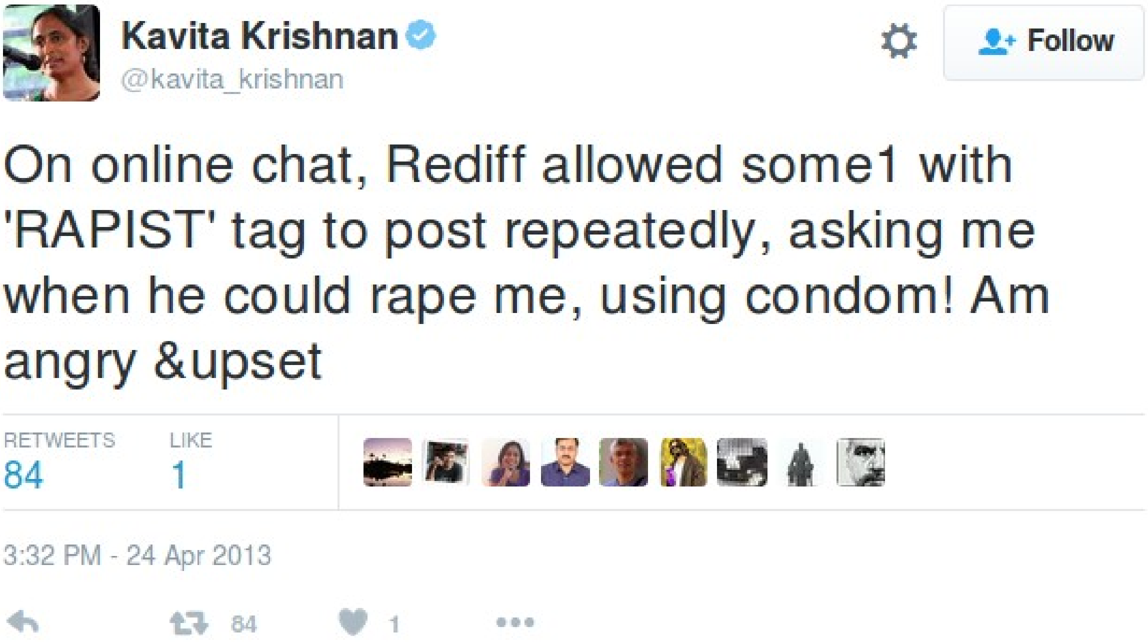
Ms. John feels that some of the abusers are emboldened by the belief that they wont ever be held accountable under law - they consider themselves as merely doing their part to promote their religious convictions, no matter how radical. She remains convinced that the technology to identify such abusers does exist, and believes that allowing them to continue their abusive behavior will only serve to create more victims.

She also believes that the systemic abuse is, to a large extent, the product of an organized syndicate with daily “delivery targets”, and performance based incentives. The fact that some Twitter handles favorite everything she tweets shows her that her Twitter profile is being actively monitored. She further reports having been approached by a political party to tweet for them – a semi-professional offer she declined. As the conversation ended there, she cannot be certain whether these “recruits” are paid, but she nevertheless believes the primary incentive is the opportunity to do something “fun” for a political cause.

Despite all the abuse and threats Ms. John and her father have received, she maintains that the incidents have not had a silencing effect, though she cannot speak for those who live simpler, more innocent lives. Even if one is brave, there are people around who will be worried; so it is not just about bravado, according to her. Today, in a bid to ensure her physical safety, Ms. John turns off all location services on her electronic devices, and refrains from posting or being tagged in too many pictures on social networking platforms. She also does not befriend anyone online, who she hasn’t personally met offline.

3.8 Kavita Krishnan; Secretary, All India Progressive Women’s Commission

In April 2013, Kavita Krishnan, activist, feminist, and Secretary of the All India Progressive Women’s Commission was invited by Rediff to participate in a chat discussing violence against women. The online discussion started off fairly well, with Ms. Krishnan picking out and answering the questions posed one at a time. However, a little way into the chat, someone with a handle 'RAPIST' repeatedly intervened in capital letters. In one 'question' he said, “Kavita tell women not to wear revealing clothes then we will not rape them.” The same man then posted another question several times: “Kavita tell me where I should come and rape you using condom.” Both questions were in block capitals and very visible. Though Rediff officials initially assured Ms. Krishnan that a FIR would be filed in connection with the incident, this was never done and all she received were a handful of questionably vague explanations including how live chats could not be screened and how Rediff officials failed to notice the abusive user since there were “so many questions”.



Kavita Krishnan
@kavita_krishnan

On online chat, Rediff allowed some1 with 'RAPIST' tag to post repeatedly, asking me when he could rape me, using condom! Am angry &upset

RETWEETS 84 LIKE 1

3:32 PM - 24 Apr 2013

[*Ms. Krishnan's tweet immediately following the attack*]

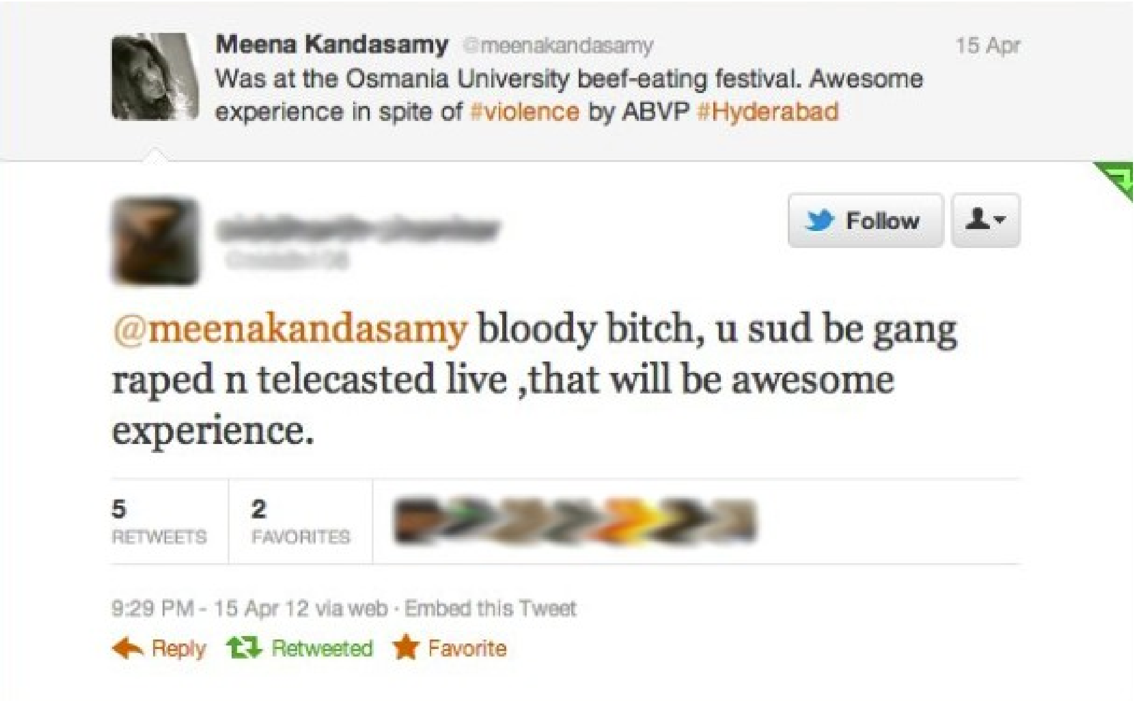
In retrospect, Ms. Krishnan tells us that this was only one of the early instances of online harassment as things have become a lot worse since she started spending more time on Twitter. She remarked that on Twitter, there are people subjecting you to hate speech, sexual intimidation and sexist comments on an almost daily basis. Ms. Krishnan receives so many violent threats over social media these days that she no longer bothers reporting them to the police unless they get very specific. More than the threats themselves, she finds the abusive atmosphere that they create to be the true challenge. She finds it difficult to keep the mind steady and cool in such a scenario and to focus on what she would like to say is not easy. It does things to your head, your sense of mental poise; it also affects your physical sense well-being, is how she describes it.

Ms. Krishnan spoke of three kinds of online abusers that she has encountered. First, there are the organized right-wing supporters, who can be identified as organized as they retweet the same tweets and generally behave in a similar fashion. Such individuals, according to her, harbor crass thoughts on sexual violence, and she treats their tweets and thoughts as expositions of their politics. The second category of organized, aggressive abusers are the so called men’s rights activists. She considers them a strange set of people, who almost constitute a subset of the first category of abusers. They usually have right wing sympathies and employ terms such as “feminazi”, though they tend to not be as aggressive as the first category of abusers and instead comment mostly on one's physical attributes. The third category, which is much smaller according to Ms. Krishnan, is not as abusive as the other two, though one may say they are aggressive. They don’t devolve to the level of abuses as such, but fervently defend their political ideologies while maintaining that your conflicting ideologies are misguided. From her experiences of engaging with such abusers, she is convinced that there are organized “abuse syndicates”, the behavioral similarities of whose members are plain to see.

As regards the reporting mechanisms instituted by social media platforms, Ms. Krishnan feels that they offer completely impersonal experiences. It’s as though there is a machine out there, which looks at complaints and makes verdicts on their legitimacy. This machine doesn’t understand regional languages, the nature of caste or gender in India, and the nature of common abuses. She finds it particularly stressful when she is told by social media platforms in the boilerplate style that the impugned messages do not violate community standards. She also finds it strange that the Cyber Cell, which is responsible for the investigation of cyber offenses is housed within the Economic Offenses wing, which admittedly has nothing to do with online offenses. Their methods of dealing with it should be much more modern and up-to-date than they are.

3.9 Meena Kandasamy; poet, writer, activist

Meena Kandasamy is a poet, writer and activist from Tamil Nadu, whose participation in a “beef festival” organized in April 2012 earned her a torrent of hateful and abusive messages on social media. While the abuses were at their peak around the time of her attendance at the beef festival, she reports having been subjected to abuse post then, and they continue to trickle in to this day.



Meena Kandasamy @meenakandasamy 15 Apr

Was at the Osmania University beef-eating festival. Awesome experience in spite of #violence by ABVP #Hyderabad

[redacted]

@meenakandasamy bloody bitch, u sud be gang raped n telecasted live ,that will be awesome experience.

5 RETWEETS 2 FAVORITES

9:29 PM - 15 Apr 12 via web · Embed this Tweet

When asked about the threats of violence she has received, Ms. Kandasamy recollects an unrelated incident, where a stranger had messaged her on social media with a request to read through and provide comments on something he had written. When she chose not to respond to this request, the stranger pointed out that he had been notified that she had seen his message, and his next line was “fuck you whore”. Ms. Kandasamy considers this incident as proof that on social media, instances of abuse can be spontaneous and not always politically motivated. For women in a public space, just the fact that they don’t give men the importance or respect they think they deserve could motivate abuse and disparagement.

Ms. Kandasamy admits that she was quite scared and frightened for her physical safety when faced with abuses following her attendance at the beef festival. Though she wasn’t immediately scared by the abuses, she feared for her life as soon as a spate of beef-consumption related murders erupted across the nation. This prompted her to approach the police and file a complaint, but no action was taken, according to her.

Ms. Kandasamy was equally unsatisfied with the reporting mechanisms offered by social media platforms. As she points out, Twitter did not have a policy against hate speech back in April 2012. It wasn’t until later, when a British journalist received rape threats over Twitter that they instituted and enforced such a policy. Her own reports before Twitter of rampant abuse leveled against her were met with responses to the effect that no action could be taken as people were merely exercising their rights to free speech. She has similar stories to tell regarding Facebook’s anti-hate speech policies, which permitted a sexist page titled “Masculinity India” to remain functional until around 400-600 people made persistent complaints against it.

Though things have recently become better on the social media reporting front, Ms. Kandasamy nevertheless feels that these platforms must have clear anti race, anti misogyny and anti-caste policies and censorship mechanisms. She also feels that there must be higher levels of cooperation between social media platforms and LEAs in their investigations into complaints of online abuse.

3.10 Navrang S B; Former Head of Social Media, BJP

Navrang S B headed the social media initiatives of BJP between 2009 and 2015. During this time, he oversaw the party's social media outreach programs under Mission 272+, an ambitious campaign that sought to rally voters with a view to securing a 272+ seat majority for BJP in the Lok Sabha. This campaign made extensive use of social media channels to engage volunteers and enable them to connect with BJP leaders, provide inputs to speeches, organise local activities and help the campaign at the booth level. Please note however, that the views expressed here are the personal views of Mr. Navrang, and must not be seen to represent those of the BJP itself.

Mr. Navrang uses the Internet every day to speak publicly on issues of contemporary significance. He finds the Internet immensely valuable in this regard for three related reasons. First, it allows him to communicate in real-time with a global audience, unlike geographically bound forms of traditional media like print and television. Second, it makes for an unbiased platform for public discourse as users of all ideological persuasions are given equal opportunities to speak their minds. Popular online speech platforms including social media websites and blogs are not limited in availability to particular sections of the population, but are open for all. Third, the Internet is very cost-effective when compared to traditional media, allowing users from all economic backgrounds to freely create and consume content of their choice. Mr. Navrang believes that these factors, when combined, make the Internet a particularly important vehicle of free expression in modern society.

While engaging with other users of the Internet, Mr. Navrang makes it a point to remain acutely aware of the sheer variety of conflicting sensibilities that it hosts, and tries his best to not make any controversial statements that might be hurtful to others. He understands that the possibility for anonymity on the Internet motivates people to say and do things online that they would never say or do in the real-world, making it very easy to trigger aggression and hostility. At the same time, he also appreciates the value that anonymity adds to online discourses, as he finds that anonymity also empowers users to openly engage in legitimate criticism in ways they never would have done offline.

Mr. Navrang does not recall many instances where he was at the receiving end of online harassment, save for one occasion. When the BJP headquarters in New Delhi was attacked by supporters of the Aam Aadmi Party (AAP) in 2014, Mr. Navrang, who was present in the attacked building and witnessed the violence first hand, sent out a couple of tweets over the following days in which he decried the attack and made some negative remarks about AAP in the process. In response, a number of hateful tweets were sent his way, most of which were aggressively vocal about their displeasure at his remarks. According to Mr. Navrang, he never tried muting, blocking, or reporting any users during this exchange, choosing instead to attempt justifying himself as to why he said what he did – pointing out that it was the pain of seeing his colleagues' injuries that made him lash out. Moreover, he has not felt silenced in any way in the wake of this incident, as he sees the whole thing as no more than an aggressive expression of differences in opinion.

Though Mr. Navrang himself has not had much experience being at the receiving end of online harassment, during his time as head of social media at BJP, he has helped numerous individuals who have faced such attacks. He recalls accompanying more than one such individual to the Cyber Cell division of the Delhi Police, who in his opinion were more than helpful in identifying the perpetrators of such attacks and holding them accountable. Speaking on the efficacy of LEAs in handling cyber grievances such as those related to online harassment, Mr. Navrang finds that approaching the right agency makes a world of difference as to how complaints are handled. While officers at non-specialized police stations may not be very well equipped in handling such complaints, Cyber Cells that are present in all major Indian cities are much better prepared for the task. He still recognizes this as a problem, seeing how it is not feasible for residents of smaller towns and villages to approach Cyber Cells. To remedy this state of affairs, Mr. Navrang recommends the initiation of focused capacity building exercises for non-specialized police officers so as to better equip them to assist citizens who approach them with cyber grievances.

On the ideal means to curb online harassment on a large scale, Mr. Navrang advocates an approach that combines both State and non-State regulation. Though the content reporting mechanisms offered by online speech platforms already help limit unwanted content to a large extent, this system is weighed down by high-volumes of reported content, which translates to long turn-around times when it comes to acting on reports received. This calls for a re-visit of content moderation practices adopted by these platforms so as to strengthen back-end operations and shorten turn-around times. As for applicable laws, Mr. Navrang finds that Indian statutes already contain an adequate number of provisions with regard to penalizing instances of online harassment. He also cautions against strengthening these laws as this could result in excessive limitations on free expression. He is of the opinion that expressions of dissent and displeasure, even those aggressively worded, must not qualify for penal sanction. It is only when such expression threatens bodily harm that the legal machinery must be invoked so as to prevent the possibility of real-world violence.

Mr. Navrang feels that the boundaries between the physical and virtual worlds are fast blurring. As more and more activities such as banking, commerce, and governance grow to depend on the Internet as a platform to function, users should divest themselves of the notion that the Internet is separate and distinct from the real-world and that the things they say and do online do not come with consequences. As the prevailing norms of social conduct would prevent most from abusing others with abandon in real-life, the same norms should apply on the Internet as well, says Mr. Navrang. It is only with such a fundamental shift in attitude that online harassment can be truly curtailed and the integrity of the Internet as a speech platform preserved.

3.11 Prabir Purkayastha; Editor, Newsclick

Prabir Purkayastha is an engineer and a science activist in the power, telecom and software sectors. He is one of the founding members of the *Delhi Science Forum* and serves on the editorial board of Newsclick – a Delhi-based viewer funded news channel. He has also written extensively on a number of science and technology policy issues. Mr. Purkayastha uses social media sparingly. He publishes some articles on his own website. Comments are moderated by him, and not many of the comments manage to get through the moderation process due to paucity of time and how much he can handle in that limited time. His website has a limited presence on Facebook and Twitter. He also writes articles for other outlets such as The Hindu, and such online publications are where he mainly sees trolling in comments.

He says that he is familiar with what happens to those who are active as he has observed it in the kind of comments people receive. He has read about, as well as analyzed the situation, but he is not a big target as he is not very active on social media himself. He does not focus on interacting on social media too much because he believes that to have an impact on social media, you have to be engaged with it in a big way and there’s no point in having a small footprint on it.

According to Mr. Purkayastha, online media can address certain kinds of niches that traditional media cannot focus on, and it breaks the monopoly of news in any part of the world. News in traditional media would be controlled by a handful of corporations that behave in a conformist manner and are controlled in different ways by the government or corporate houses. Topics like poverty are reported more in the form of statistics than journalistic human interest stories in traditional media, because advertisement revenue requires relatively positive news, as people who feel good would buy more. Online media provides and alternative to this, but also brings extreme writings with it.

When asked about censorship, Mr. Purkayastha said that freedom of expression can be curbed in two different ways: (a) it can be curbed by government agencies when sedition charges are filed, and (b) mob censorship through trolls whenever you write anything critical of the Government policies, taking the argument away from actual criticism of the author's or journalist's writing. He mentioned that women are particularly vulnerable to mob censorship, as they get vicious sexist abuses in addition to the off-topic political criticisms and comments that other writers face online.

Regarding anonymity, Mr. Purkayastha said that though it is important, it should not be a complete cloak for every individual. There is a need for a legal process through which anonymity can be revoked if someone is being trolled beyond a certain point. He also spoke about the need to build a culture against trolling and making it an unacceptable activity as trolling on social media takes place in a public space, and the norms of behavior in a public space should apply. He does not see anonymity itself, but rather a certain kind of culture and group mentality as the problem. Anonymity plays a role, however most people are happy to put up their pictures on Facebook and to identify themselves as the ones who took certain actions. He thinks that allowing intermediaries such as the social media networks to track the activity of trolls and to block trolls themselves would give a lot of power to the intermediaries to decide what is acceptable and what is not.

Mr. Purkayastha, in conversation with SFLC.in, said that online harassment has a silencing effect on those who are sensitive and who do not want to be very visible. Giving the example of Shobha De, he said that if you're already into it and won't be affected by it, then you won't be attacked any way. Particularly for women, though, there's a limit to it. A lot of women have left the online space because of the continuous trolling that they receive.

He called it the responsibility of everyone who recognizes this problem to promote writings which expose this culture. His solution for this problem involves naming and shaming the trolls, so that eventually the culture of trolling is no longer acceptable, the way that spitting "paan" on roads and walls of buildings was once very common in India, but has now become an unacceptable behavior. Such change in the nature of the community as a whole is a gradual effort and takes time, but it doesn't happen by itself. Groups are needed to take this up not only as a defensive activity, but also as a preventive activity, with naming and shaming considered as a preventive activity by Mr. Purkayastha. Taking a few prominent cases to Court can also act as a deterrent for others. Mr. Purkayastha suggested creating a body or a group of people who can put forward the resources needed to take up such matters on behalf of individuals.

3.12 Rajeev Chandrasekhar; Member of Parliament, Rajya Sabha

Mr. Rajeev Chandrasekhar is an independent member of Parliament, currently serving in the Rajya Sabha from the state of Karnataka. He actively uses social media platforms like Facebook and Twitter to interact with citizens of the country, and in his opinion these online platforms are set apart from the traditional media due to their capability to effectively communicate with a large audience that traverses geographical boundaries while cutting costs. Being a politician, he has often faced remarks and comments that are abusive and meant to harass him, but Mr. Chandrasekhar maintains that he is not thrown off his stride by trolls or such similar behavior.

Mr. Chandrasekhar opined that the Internet is the largest audience ever gathered, and when an idea is tossed in the cyber world, the discourse is automatically expanded to a larger audience that is bound to attract all types of people. On the contrary, in the real world, if one conducts a discussion on an offline platform, due to the time, effort, and other geographical boundaries attached to being physically present, only the select few who are genuinely interested will make themselves available for an event. However, he maintains that a certain amount of unruly behavior and harassing speech is the cost of being able to get a large audience that is quicker to assemble and cost effective.

Mr. Chandrasekhar nuanced his stance by stating that the ease with which people behave online, and would probably not indulge in such behavior offline, can be attributed to the theory of being in a crowd. According to him, amongst a tightly knit small gathering, an individual will be self conscious, in the offline or online forums. However, if the audience is increased multi-fold even in the physical world, a person might tend to be less careful about their actions and words. In his words, “crowds give you a cloak to be things that you otherwise won't be, but that is normal human behavior - it has nothing to do with the cyber space or with the Internet.” However, as tools become available to discover the identity of a person, even in a mob in the physical world, for example, with the help of CCTV footage, the garb of anonymity feels less secure. Per Mr. Chandrasekhar, the same approach will be applicable in the online world as it evolves mechanisms to clearly identify users in case a breach of law and order occurs.

Talking about his responses to abusive and harassing comments and messages targeted at him, he labels them as the part of the discourse of addressing a large audience. Personally, Mr. Chandrasekhar's usage or engagement in various platforms is not deterred by being at the receiving end of abusive speech. However, he does exclaim that even if this is the cost of using these platforms, the intermediaries are liable to ensure that content that is in contradiction to law and order, for example ISIS and terrorist propaganda is removed swiftly. As a suggestion for content moderation, Mr. Chandrasekhar believes that manual intervention with respect to disabling accounts or blocking Twitter handles is not going to be the solution owing to the vast volumes of content generated on these platforms on a daily basis. He strongly suggests development of automated mechanisms.

With respect to the approach to be taken at the regulatory, user, or legal level to tackle this issue, Mr. Chandrasekhar opined that rules, and laws should be rolled out simultaneously at all levels to ensure a smooth functioning of the entire ecosystem wherein everyone is aware of their roles and responsibilities. Specifically for LEAs, he mentioned that a culture of safe and transparent implementation of law in the cyber world would assure the netizens of the capability of LEAs to maintain an environment where free speech is not muzzled and subsequently, with law as a deterrent, abusive behavior will phase itself out in due course.

On the capacity building of police force and the LEAs, Mr. Chandrasekhar however, did admit that sensitivity towards online abuse and harassment is negligible where such behavior is looked at as an “elitist little play thing” and in his opinion, to change such perspective, the government, along with the law enforcement needs to internalize that the cyber space is simply an expansion of the real world where one requires laws, their strict implementation, and adjudication by the judiciary.

As practical suggestions and improvements on parts of platforms, Mr. Chandrasekhar suggested creative measures such as being able to block and mute one person and all his followers on a platform like Twitter. His justification behind this being that it is quite plausible that an aggressively abusive Twitter handle would be followed by many of those who tend to create an army of trolls and launch an attack. In this way, a user is able to shield himself from this entire ecosystem. Another interesting suggestion that Mr. Chandrasekhar made for tools that can be used by the platforms is to keep a check on the accounts that are regularly reported and enable a feature wherein a user can block all accounts that have been previously blocked by a particular user or by anyone on Twitter. With measures such as these, he believes that users gets further customization tools at their disposal and also the choice to have the entire Internet as their audience or limit it by certain filters.

On the viability of one of the frequently used measures of counter speech by various platforms, Mr. Chandrasekhar disagreed on it being a successful tool. As per him, injecting a reverse narrative would end up being a pointless blame game, which would just snowball into a competition over which side gets exhausted first and gives up.

3.13 Rakshit Tandon; Cyber Security Expert and Consultant

Rakshit Tandon is a renown cyber security expert and consultant, who is also a hacker. He has been assisting various LEAs in investigating cyber crimes since 2008, and he conducts routine training sessions for the officers in this area. He also conducts numerous seminars and lectures at schools and universities so as to better prepare the youth to be well-rounded digital citizens, and secure themselves from online sources of harm. Mr. Tandon very frequently uses the Internet as a communication medium. He finds great value in social media platforms, and considers the ability to create user groups on services such as WhatsApp to be specially significant as this enables him to easily exchange information with a large audience.

Mr. Tandon receives roughly thirty to forty complaints each day related to abusive uses of the Internet. A number of these complaints are made by individuals who have had fake profiles set up in their names on social media platforms, and others are from users who have their personal content like photographs and contact numbers posted on various public forums. An increasing number of complaints also come from male complainants, which as Mr. Tandon points out, is indicative of how the underlying problem is more or less gender-neutral. Setting up fake profiles of powerful political figures and using these profiles to spread rumors and misinformation is also a frequent source of complaints. Mr. Tandon reports offices of the Uttar Pradesh and Haryana Police, with whom he frequently collaborates as part of his work, as receiving four to five such complaints every fortnight.

Mr. Tandon aids LEAs in their investigation of such complaints on occasion, including in determining the legality of particular content. He finds that law enforcement officials are not immune to misjudgments as to the legality of content. This makes it essential to take special care in classifying content as illegal or permissible, lest legitimate free speech be restricted in the process. Once any content is identified as illegal and deserving punitive action, the next step is to trace its creators who on many occasions will be well-hidden behind IP addresses and proxies. As the process of law must mandatorily be followed in this exercise, this involves approaching relevant service providers like Facebook and Twitter and requesting user information. Aside from procedural delays, rampant use of fake identities and SIM cards are factors that stand in the way of identifying perpetrators and bringing them to justice in a timely manner. Methods of traditional policing including ground-level investigations are used to tackle these challenges to the extent possible.

On the LEAs' state of preparedness when it comes to registering and investigating cyber crimes including harassment, Mr. Tandon admits that there definitely is a problem, though he also believes the situation is slowly but steadily improving. While many law enforcement officials remain apprehensive of taking up such complaints, the efforts by Mr. Tandon himself and many others like him in training and empowering them to handle the complaints have definitely yielded results.

Since 2009, Mr. Tandon has been traveling the nation, and educating members of Central and State Police Academies on effective investigation protocols, booking offenders and gathering physical evidence to support the charges filed. He also trains the officers to gather information from publicly available information pools such as social media websites and leverage this information towards proactive policing.

In most cases, when individuals approach him for assistance in relation to online harassment, Mr. Tandon observes that they are not as interested in identifying their harassers as they are in getting the harassment to stop. In this regard, Mr. Tandon helps them get in touch with the concerned service provider by way of existing content report mechanisms, including “report abuse” buttons built into social media platforms. He finds that content such as sexually explicit images and pornography, which patently violate applicable content policies, are almost immediately taken down. As he rightly points out, while these steps can easily be taken by the complainants themselves, the fact that they continue to approach him nevertheless is demonstrative of the low levels of awareness amongst users as to the available channels for reporting and removing harassing content online.

According to him, most individuals who approach him are entirely unaware of even basic concepts such as security and privacy settings, geo-tagging and location data, content visibility settings, and malicious online content. Mr. Tandon observes a lack of awareness amongst the harassers as much as the harassed, in that they are blissfully unaware of applicable laws and content policies. When arrested, they claim that it was all said and done purely for fun, and that they did not anticipate the existence of laws prohibiting such conduct. Mr. Tandon registers his amazement at coming across even highly educated people who claim to be unaware of laws against harassment. He feels this lack of awareness is aided to an extent by the myth of anonymity, which leads people to believe that nobody watches what they say and do on the Internet. The habit of forwarding damaging messages without so much as reading through them first was also cited as a contributing factor.

All the above being said, Mr. Tandon is of the opinion that a portion of the blame lies with Internet intermediaries serving as speech platforms. Even if major players like Facebook and Twitter are US based companies, as they cater to a potential customer base of 1.3 billion Indians, they need to mold their content policies to fit local religions and cultures. He feels that the platforms do not take into account the multiplicity of Indian languages, which means a number of content policy violations made in local dialects are overlooked. The response time with respect to reported content is yet another area of concern, and must be improved. Lastly, Mr. Tandon believes the platforms must make more efforts to educate their users on how to use their services, including content reporting tools.

3.14 Ravish Kumar; News Anchor, NDTV India

Ravish Kumar is one of the most respected Hindi news anchors in India, who in August 2015, made a public exit from social media platforms including Facebook and Twitter. Known for his unbiased coverage of Indian news and polity, Mr. Kumar had previously been very vocal on social media and was an active presence, particularly on Twitter. In an interview with Scroll.in, an online news website, he famously said, “I have stopped tweeting because social media space is no longer a citizen’s space. It has been usurped by political parties to peddle their ideology and propaganda. It’s an online lynch mob where anyone with organizational support of 500 can send out 10 lakh tweets and declare me a thief.”

In conversation with SFLC.in regarding his Twitter exit, Mr. Kumar said that he used to receive extremely violent threats and abuse at regular intervals, which started affecting his role as an anchor. He felt it was his duty as a journalist to be unbiased, and he could not continue using Twitter as an individual with his own independent opinions. He said that although Twitter had been very helpful in providing a multiplicity of voices, but rampant and targeted abuse made it very difficult to use the platform effectively.

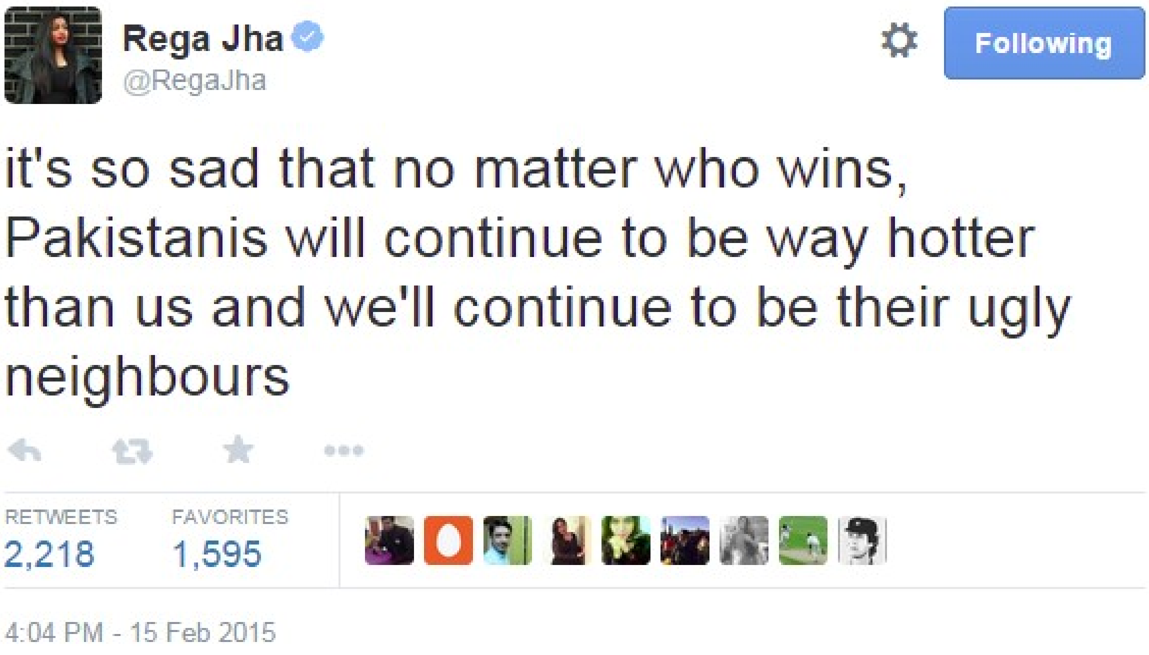
On the differences between interacting with people in person as against over social media, Ravish felt that the same people, who are more than friendly in person, turn aggressive and resort to violent abuse over social media. People cease to be social organisms on Twitter, said Mr. Kumar. He felt that the more he saw society, the more convinced he was that people are not real pallbearers of democracy. Even the idea of democracy is an illusion according to him, who sees it more as a business where people are divided on the basis of caste and religion.

He then recounted his experience of approaching the Cyber Cell (the cyber-crime wing of the state police) to file a complaint when his website was hacked. He was forced to withdraw his complaint as the police were unaware of what hacking even meant, and was unsure of how to handle such a complaint. He also feels that blocking and reporting abusers on social media are not effective remedies, as the sheer volume of abuse makes this impractical. The fact that there are fake as well as genuine accounts that indulge in abusive behavior only exacerbates the problem. Mr. Kumar suggested user verification as a potential means to bring some accountability to social media use, and thereby curb its rampant abuse.

Today, Mr. Kumar has become somewhat of a social media pessimist, who feels that there are only illusions of freedom of speech and multiplicity of opinions on social media, and that they have ceased to be social media in the true sense of the word.

3.15 Rega Jha; Editor, BuzzFeed India

On 15 February 2015, while a much anticipated cricket match between India and Pakistan was in progress, Rega Jha, Editor of BuzzFeed India, made a quick tweet that read, “it's so sad that no matter who wins, Pakistanis will continue to be way hotter than us and we'll continue to be their ugly neighbours”.



Rega Jha
@RegaJha

it's so sad that no matter who wins, Pakistanis will continue to be way hotter than us and we'll continue to be their ugly neighbours

RETWEETS 2,218 FAVORITES 1,595

4:04 PM - 15 Feb 2015

Jokes surrounding the relative attractiveness of Pakistanis as against Indians are by no means new, and in Ms. Jha's own words, about 30 seconds of thought went into her tweet. It was meant to be nothing more than a humorous quip, which on any other day would likely have been forgotten after a few re-tweets and likes. However, as the tweet was made by a relatively visible figure during a politically charged India-Pakistan cricket match, when millions of Indians were glued to their television screens fervently hoping for India’s victory, it was seen as an insult by many.

Soon after the tweet was posted, Ms. Jha was bombarded with a torrent of angry and viciously disparaging responses, some going so far as to suggest that she deserved to be raped and dragged around in public. Facsimile accounts bearing her name were set up to make periodic self-derogatory tweets, and even a few celebrities joined the fray, expressing their displeasure at her tweet (sans the threats of violence and rape). Within a matter of hours, the name “Rega Jha” was trending on Twitter, besting hashtags such as #BleedBlue and #IndvsPak associated with the match itself.

Though Ms. Jha quickly tweeted an apology and retracted her offending statement, the abuses continued to pour in for days afterwards.

Recalling the incident, she said there were only a handful of tweets that made direct threats. A lot more were very sexist jokes and insults, on the lines of “you deserve to be raped” rather than “I am going to rape you”. A particularly vile tweet that stuck with her was something along the lines of “you deserve to be put in a burkha, raped and dragged around Saudi Arabia”. Moreover, she was baffled that men could make the same joke on Twitter and get away with it. Men had been making that joke for years, she said. The backlash in her case, she believed, was due to the fact that a young, liberal woman had never made the joke so outspokenly.

She further speculated that she might already have been on the radar of those politically affiliated to the far right. As her job involved making jokes on India’s pop-culture icons, she routinely made jokes surrounding Prime Minister Narendra Modi, who is as much a pop-culture icon as he is a politician. Ms. Jha believed that the frivolity with which she had been making such jokes might have drawn considerable ire from the right-wing loyalists and contributed to the incident.

On being asked about how she dealt with the abusive messages and threats, she said that her first response had been to take screenshots of the abuse, as such messages are often deleted as soon as their authors are called out. She went to the Deputy Commissioner of Police the following day, showed him the screenshots, and told him she was being threatened – that she felt unsafe. He had already heard about the incident and was apologetic of what had happened. He also said that the police received several complaints from women every month about online harassment, but they are not equipped to do anything about them. As per Ms. Jha, the DCP seemed very much like he wanted to help, but did not know how.

Today, Ms. Jha does not hesitate to admit that the whole incident had a perceptible silencing effect on her. Terming it a side-effect of being a liberal, outspoken woman on Twitter in India, she said it takes a lot of energy to not get silenced. When for every trivial joke, there are people calling you anti-national, bitch, slut, and suggesting that you be raped, dragged around and the like, it stops being fun, she said. So, even if she had not made a conscious decision to tweet less frequently or to tweet less outspokenly, it naturally just happened that she used Twitter less.

As regards the measures that might help curb the possibility of such abuse of online speech platforms, and the platforms’ own responsibilities in this context, Ms. Jha wasn’t sure if the platforms were to be blamed. According to her, it all comes down to the fact that humans are horrific when allowed to be anonymous and when their ability to empathize is taken away. When there aren’t faces and voices attached to people, it is very easy to insult them, and that isn’t the fault of the platform necessarily, she said. In her experience, both Facebook and Twitter were extremely responsive - when harassment complaints were filed, Twitter specifically was prompt in disabling the responsible accounts. As far as the platforms are concerned, much of their business is sustained on large scale influences, and large scale influences are the ones that get all these abuse and insults. As the abusive behavior is driving large influences off the platforms, the platforms themselves are threatened, said Ms. Jha. The last thing they want in such a scenario is for their users to be silent. The last thing they want is for us to be quiet, she said. She thinks online platforms are doing whatever they can to fix the problem, but their capabilities in this regard are limited as they live under the shadow of issues such as freedom of speech and censorship. According to Ms. Jha, the problem with the system is human nature, which is also what makes it unfixable.

3.16 Rohit Chopra; Associate Professor, Santa Clara University

Rohit Chopra is an Associate Professor of Communication at Santa Clara University and has been Visiting Scholar at the Center for South Asia at Stanford University. His research addresses the relationship between media and culture, new media technologies, and how media shapes political and cultural memory. He is also well-known as the man behind the popular Twitter parody account @IndiaExplained. Unlike the other interviewees, Mr. Chopra does not live in India. He lives and works in U.S.A. and his experiences with resolving issues related to online harassment reflect on the differences in the approach to online harassment in the LEAs of India and the US.

Mr. Chopra first started using online forums in or around 2001. At that point of time, online hate speech was limited to the confines of specific discussion threads, and did not follow people outside of the forum of discussion. He observed hostility or animosity in the way Indians and Pakistanis would converse about cricket, and he could sense misogyny when occasional misogynistic or chauvinistic remarks like ‘You’re a woman, what do you know about cricket?’ were targeted at women and people whose handles looked like they were women. He later wrote articles on another forum dedicated to social/progressive issues, where some of the responses were “fairly nasty” but still milder than the kinds of messages he has seen on social media. He has also used Twitter for about 4 years with two different accounts: one on his own name, and one as a parody account called ‘Rushdie Explains’ now 'India Explained'. He received negative remarks from conservative or reactionary Muslims on the parody account, which he said may be because, despite a clear parody label, they did not understand that it was a parody account. At about the same time, he started to face a systematic attack from the Hindu right because he had started to comment on Prime Minister Modi in a humorous vein while trying to make a sharp political point making him a target of both

Hindus and Muslims.

Mr. Chopra has faced about four sustained bouts of online harassment on the parody Twitter account, out of which his university was targeted on two occasions and he was personally attacked relentlessly on the other two occasions. According to him, people are not targeted only for what they say, but anyone with any sort of prominence is usually targeted, even if they've made a fairly innocuous remark about the troll's favorite politician. He discussed the existence of a notion in the Indian society of only certain people having the right to talk about certain topics. He considers trolling to be a universal phenomenon that is not limited to India.

Mr. Chopra referred to the existence of two different kinds of trolling on the internet: (1) where women are targeted, such as the Gamer Gate scandal in which Anita Sarkeesian was targeted for drawing attention to the objectification of women in gaming, and (2) politically motivated trolling, Russians, he said, have been very adept at spreading misinformation. He also remarked upon the speed and efficiency with which such trolls react.

Mr. Chopra called anonymity a double-edged sword. He mentioned the importance of anonymity in protecting activists in places like Saudi Arabia and India, and the need of anonymity for whistle-blowers. He gave an example of the need for offline anonymity, saying that the act of making RTI applications publicly available puts the lives of activists in danger, and effectively eviscerates the right to information. On the other hand, he said that anonymity creates an impunity. Here he mentioned how several newspapers had (or had considered to) shut down their comments section because the comments turn into a place of vile arguments. He thinks that many of the people who engage in abusive online behavior do not take the effort to hide their identity beyond a certain point, such as by using a VPN or other methods of masking themselves. Thus, it is possible to track down their identities.

With regard to the responsibility of social media platforms, Mr. Chopra said that they need to take more action, rather than deferring all responsibility as a neutral meeting place. The platforms need to be proactive when there is abusive behavior. Mr. Chopra's experience with Twitter's complaint mechanism has been less than ideal. His complaints on serious threats and abuses were not even responded to. In one case, someone had tweeted that they would rid Mr. Chopra of the US soil. Twitter's response was that they did not feel that it rose to the level of abuse. In this case, even the police in the US had told Mr. Chopra that it was a clear threat. In his opinion, Twitter only acts when its own brand is at stake, such as when a celebrity is involved or people are engaging in copyright violation. He said that as far as he is aware, it is not possible to even describe the situation while reporting something on Twitter. They have a standard form, and it is impossible to get to speak with a human.

He suggested implementing more nuanced, granular and more differentiated reporting mechanisms, a very clear statement of what is considered as harassment under Twitter's policies, and a structure of accountability, with instructions on what to do if someone feels harassed. According to Mr. Chopra Twitter hides behind the First Amendment, but they can have certain policies for inclusiveness and making sure people are safe and are not threatened without going afoul of the First Amendment. Additionally, Mr. Chopra mentioned that the standard of the First Amendment was being followed by Twitter worldwide, but there's no reason for the First Amendment to be the operating principle in Indian discourse.

His experience with law enforcement in the US was the opposite of most of the responses received by our other interviewees from law enforcement in India. He said that he had reported two incidents in the US, and the law enforcement had monitored them and kept them in perspective.

Regarding the silencing effect of online harassment, Mr. Chopra said that two things happen: (1) a silencing effect, and (2) a signaling effect, also known as the chilling effect. The chilling effect is used by corporations as well. Here, he gave an example of a $5 million lawsuit that had been brought against an old lady for downloading some songs. That lady could not pay such an amount of money, but the idea behind the case was to terrify and deter college students from downloading songs.

He talked about a study being done on Twitter by a Sangeet Kumar, who had mentioned that less than 2% of Indians are on Twitter, but a lot of them are seen as influential people. Even with its minuscule reach in percentage terms, it's an influential medium, and Mr. Chopra considers it extremely harmful for public discourse that people are intimidated by it.

When asked for his suggestions on solutions for the issue of online harassment, Mr. Chopra said that there are two ways to approach the problem:

(1) There are specific things that each platform can do, and platforms will not do these things by themselves. Some kind of regulatory measures could be taken. There could be some kind of flexible general principles. There can be legislation which applies to specific platforms, but the way to do this would be to look at it from the perspective of harm and the principle of being inclusive.

(2) Bringing about a change in the culture, and re-establishing the credibility of online journalism. Mechanisms for flagging a rumor or potentially abusive content would help.

3.17 Saikat Datta; Journalist

Saikat Datta has been a journalist for over 19 years as an editor and an investigative reporter with several news organizations. He has been a defense correspondent with *The Indian Express*, an assistant editor with the *Outlook* magazine, Resident Editor with *DNA*, member of the Editorial Board with Zee News and Editor (National Security) with *Hindustan Times*. Being a journalist, Mr. Datta uses online speech platforms like social media websites on an hourly basis for various needs such as understanding audience metrics through Twitter or political opinions on Facebook. He also writes professionally on LinkedIn about issues related to intelligence and security. He believes that digital media has its own advantages of traditional media in terms of how in a digitized environment, one can have multiple levels of communication and this has democratized the flow of information.

When it comes to anonymity and the paradox of anonymity, Mr. Datta is not in favor of it for two reasons. He feels that firstly, it encourages online abuse, as people take advantage of anonymity to express their views in very virulent and vicious ways. Secondly, the moment we start acknowledging anonymity, we in a way are also acknowledging that the environment around us is not conducive of free flow of ideas, which is a much bigger problem. Mr. Datta feels that this needs to be addressed, rather than just looking at anonymity as a means to express oneself.

Mr. Datta has been facing online harassment on a daily basis, and once was even given a death threat. However, he does not take such incidents too seriously, as he feels that they come with the profession. If people have strong views about writings and wish to let out some angst, Mr. Datta believes they should be allowed to do so. Consequently, he mostly tries to ignore such incidents and as far as possible to engage people. He also points out that when one starts politely engaging with people, most of the abusers end up appreciating their mistake. He also recounts having made friends with such abusers, as they realized that he does not have personal agendas, but just a point of view that he would like to hold on to. But most other times, he ignores harassment, and only in very rare circumstances has he ever blocked anyone online.

Mr.. Datta has rarely tried reporting online harassment to law enforcement because he doesn’t feel law enforcement is the solution to such issues. According to him, this would introduce an instrument of the state to curb speech and this can lead to undesirable consequences. He therefore feels that law enforcement must have a minimal role in matters of speech. On the limited occasions when Mr. Datta has tried reporting such incidents with online platforms, his experience was mixed. He finds Facebook very disappointing in this regard, but Twitter was seen to be much better.

Mr. Datta points out that online harassment can have a chilling effect on people, and he has in fact met many prominent people who have gotten off the platforms because of online abuse. They complain that the reasons that they had come to social media in the first place, i.e. to have conversations, make connections, understand each others point of view, have now disappeared and replaced mostly with abuse and violence.

While Mr. Datta does support counter speech a viable limiter of online harassment, he finds support groups much more valuable. He feels that support groups that work together as a network are very important, but the idea of intermediaries as support groups does not resonate with him. He doesn't find this very practical because these are profit driven private entities and he doubts how far they would go from an adequacy point of view.

As regards the shortcomings within the present system of content reporting offered by social media platforms, Mr. Datta feels that the mechanism used by Facebook in particular is very inadequate to deal with online harassment and hate speech. He has observed incidents, where violence escalated and actually penetrated the community, which is very dangerous. At that point where violence starts spiraling out of control, it becomes violence at specific level, violence against some marginalized community, against gender, against minority. This could prove to be a very dangerous situation and it becomes impossible, even for people who had started the whole thing, to maintain control.

Mr. Datta feels that we need to deploy technology in much smarter ways to try and address such issues, and notes that platforms are already doing this, though a lot more can be done. He urges platforms to make greater use of existing technology that allows them identify online harassers through IP addresses and using the content that they generate. Even from a business perspective it is very dangerous for platforms like Twitter to be unable to protect conducive free speech, as this opens them to the risk of losing people who could otherwise contribute in a meaningful manner.

3.18 Sheeba Aslam; Journalist, scholar and Islamic writer

The story of Sheeba Aslam – journalist, scholar and Islamic feminist writer – is altogether different from the others cited in this report , and illustrates the bizarre and unforeseen consequences that may befall those seeking the recourse of law from instances of online abuse.

Ms. Aslam informed us in conversation that her views expressed offline on the mismanagement of public property and funds by certain Islamic clerics had already caused her to be attacked thrice, the third leaving her home ransacked and her belongings destroyed. Later in August 2011, an anonymous user that went by the name “trueheartedindian” sent Ms. Aslam a series of ominous e-mails that warned her to stop posting anti-Indian comments on Facebook or “be prepared for consequences”, referring to what she described as ‘misogynistic traditions’ in Muslim and Hindu communities among others.

Fearing for her safety, she filed an FIR with the police and subsequent investigations were successful in identifying her online attacker as an individual who had taken unilateral offense at her Facebook comments and decided to take matters into his own hands. However, in an unprecedented turn of events, not only was her attacker acquitted of all charges by the trial court hearing her case, but the trial judge upon finding a few “objectionable” comments on her Facebook feed also ordered that charges be framed against Ms. Aslam under Sections 153 A, 153 B and 295 A of the Indian Penal Code (promoting enmity between different groups; imputations, assertions prejudicial to national integration; deliberate and malicious acts intended to outrage religious feelings).

In quashing the charges against “trueheartedindian”, the trial Judge observed: *in my considered opinion, raising national slogans, opposing anti-national sentiments and words and opposing anti-establishment words, and asking for fair debate on anti-national sentiments of any person no way attracts the criminal law even if the words used are harsh in nature and are covered under the provisions of Fundamental Rights as provided under Article 19 of the Indian Constitution.*

Ms. Aslam approached the High Court of Delhi in appeal, where the matter has been heard and is awaiting judgment as of March 2016. During the proceedings, the Judge at the High Court noted to Ms. Aslam's relief that the lower court had failed to perform its duties by being biased and becoming party to the proceeding. This notwithstanding, her experience serves as a harsh reminder that even legal functionaries such as Judges, who are tasked with providing fair hearings and making unbiased verdicts, are not infallible and may very well err in administering justice or administer it to the wrong person.

The charges against “trueheartedindian” are also likely to be dropped as the police claims to have found no evidence against the accused.

3.19 Findings

Of the 18 individuals that SFLC.in spoke with for the purposes of this report, only two admitted that their experiences with online harassment has had enough of an impact on them to force them into silence. Most others either dismissed the attacks as nothing more than the predictable misuse of liberties that come with the Internet, or continued to use social media as they did before, despite being emotionally distressed over the harassment. However, this is by no means a reflection of how online harassment affects Internet users in general, firstly because this data set is too small to paint a representative picture, and secondly because almost all interviewees, in light of the nature of their work, may be more resilient to online harassment than the average user. More detailed studies must be conducted over longer periods of time with much larger numbers of users before even a speculative analysis can be made of the effect of online harassment.

Several interviewees also speculated that there may exist organized “abuse syndicates” with explicit mandates to intimidate and subdue voices that advocated ideas contradicting their own, though none could state this to a degree of certitude as they lacked the requisite evidence.

A common theme that emerged from our conversations with stakeholders was that law enforcement officials were woefully under-prepared when it came to holding perpetrators of online harassment accountable for their actions. Police officers were reported as being uninformed on the modalities of handling Internet-related grievances, and all interviewees who have approached them with such a complaint were informed that there was nothing the police could do so far as identifying their attackers was concerned. Though some officers were inclined to help in what way they could, proffered advice in these cases were less than helpful, mostly along the lines of limiting social media use to cut the problem at the source.

On the existing mechanisms of abuse reportage and redressal available with social media platforms and the experience of utilizing them, the interviewees offered widely differing views. Some were of the opinion that the platforms were doing all they could to limit abuse on their networks, but their capabilities were limited in this regard as they were bound by censorship concerns and also the impracticality of dealing effectively with the sheer volume of grievances. Others found the reportage frameworks lacking in terms of user experience, and felt that the degree of automation and lack of actual human interaction involved in reporting abuse considerably eroded the experience. Yet others felt that the platforms were not in fact doing all that they could to limit abuse, and believed their anti-abuse policies and reportage processes could do with much revamping.

In summation, the degree to which online expressions of hate and online harassment affected particular individuals was seen to be highly subjective, depending to a large extent on the victims’ own emotional resilience and their threshold for tolerating targeted negativity. Our conversations did however highlight a pressing need for capacity building amongst law enforcement officials, with focus on effective documentation of complaints and awareness of available recourses, legal or otherwise, to online abuse. Further, there would be immense benefit in streamlining the abuse reportage mechanisms available with social media platforms by better elaboration of available options and greater transparency in the redressal process.

IV. Roundtable consultations on hateful and harassing speech online

4.1 First roundtable (New Delhi; July 28, 2016)

SFLC.in organized a roundtable discussion on 28th July, 2016 in New Delhi to initiate a focused and collaborative dialogue around the increasingly important issues of online harassment and hate speech. This roundtable was intended as the first in a series of discussions around said issues, and was attended by representatives from various stakeholder groups including intermediary platforms, civil society groups, and media houses, along with individuals who had personally experienced such online abuse and harassment. The core objective of this discussion was to recognize and understand the vast range of concerns that exist in this sphere, in an effort to develop a framework for the regulation of such activities, without stepping on the right to freedom of expression. The discussion was conducted under Chatham House rules so as to facilitate an uninhibited exchange of views.

Over the course of the event, the complex and multifaceted nature of its overarching theme unraveled, as the discussion moved from underlying social constructs, to responsibilities of intermediary platforms, adequacy of existing laws, sensitization of everyday users and effective handling of grievances by LEAs. At the very outset, it was highlighted that social media platforms, with their increasing popularity, are being considered centralized hubs for businesses and others. However, individuals, communities & institutions often find themselves at the receiving end of sustained abuse and threats either on grounds of their actual or perceived characteristics, or over their online expression. The dynamic discussion that ensued brought to light significant concerns that would require a collaborative effort across stakeholder groups to address. For the sake of clarity, we are categorizing these learnings under the following heads:

- **Conceptual understanding of online harassment and hate speech:** It was discussed at length that hate speech and speech that culminates in harassment on the online sphere, are reflective of the social outlook of the country at large. Women were seen as more frequent targets of harassment in the form of rape threats, sexual remarks, and name calling, whereas men are mostly called out for their beliefs and opinions. When discussing hate speech relations, it was considered important to take note of the power dynamics at play amongst the stronger groups, and the vulnerable ones. Limiting such content gets specially complicated considering the apprehension that in an effort to monitor hate speech and harassment, free speech may get stifled. The paradox of anonymity being an enabler of free speech, as well the reason for unabashed harassment adds yet another layer of complexity to the issue. Moreover, it was felt that a nuanced distinction needed to be made regarding the systematic attacks by online mobs against a particular person, as opposed to hateful and/or harassing speech that engages on a one to one level. This all culminated in a realization that this issue goes beyond the online domain, into the societal mindset that is amplified on the Internet, and that the faint line between free speech, and hateful and harassing speech is very difficult to pin-point.

- **Role of intermediaries:** It was the opinion of the representatives of intermediary platforms at the roundtable that the current legal frameworks in the country are sufficient to tackle this issue and they should operate in compliance with such laws. While the specific terms of service may differ in terms of permissible content depending on the type of service being provided by the intermediary, these platforms do invariably keep a check on the content being generated and evaluate it for compliance with the applicable terms of service. Additionally, platforms that have the option of users creating and generating their own content, give the user various tools such as block, filter, un-follow, and other customized options to moderate the content they receive. Though the intermediaries, in their own words 'are not a delete squad, but a compliance team', it was said that they run the perpetual risk of either censoring content that should not have been censored, or not censoring enough of the content that should have been censored. This incentivizes them to exercise zero-tolerance policies in certain areas such as child sexual abuse or terrorism, and resort to immediate take down of content related to such themes. However, in spite of the sheer volume of material that is generated and reported, it was felt that a completely automated approach cannot be followed for filtering hateful and harassing content that violates terms of service. Taking down content requires processing various factors that determine the context of that material, and this calls for a subjective approach that can be done only by a set of human eyes. Therefore, the intermediaries do have some tools for users that protect them from hate and harassing speech, and they work with certain safety experts to ensure that the users feel safe while using their services but there is always room for improvement.

- **Adequacy of legal frameworks:** A distinction was drawn over the course of the discussion between hate speech as a social as opposed to a legal concept. For legal purposes, speech would not attract penalties until it incites a real threat of violence and civic disorder. However, the law is not sufficiently equipped to deal with speech that does not incite violence, but causes psychological damage. It was undisputed that the concerns in this area cannot be solved by creating more statutes. Going down this road could lead to the creation of an equivalent of the now-repealed Section 66A of the Information Technology Act, 2000 that would lead to censorship through law and cause a chilling effect on freedom of expression. It was emphasized that the existing laws have adequate provisions, but a strict implementation is required.

- **Response from LEAs:** An evaluation of this point led to the conclusion that people who are harassed online, or are the targets of hate speech, are hesitant to approach the police and LEAs for their help. There have been instances where the police is unable to help due to the limited application of laws in such cases, as mentioned above.
- **Possible remedies:** As a part of this roundtable, SFLC.in had proposed a set of best practices aimed at limiting hateful and harassing content online. These were intended as self-regulatory measures that could be followed by intermediaries functioning as speech platforms, where users could create and publish content without pre-filtrations. Amongst the measures that were discussed extensively was the practice of promoting ‘counter speech’ on the platforms that are most frequently used to spread hateful propaganda and harassment. This was generally seen as an effective counter-measure deserving further exploration, and one of the intermediaries mentioned a project they were formulating on ‘counter radicalization’. However, concerns were raised with respect to the identification of areas that would benefit from counter speech, and its effectiveness with respect to mob attacks. Another unique approach suggested by the participants was to ‘vaccinate’ first time users by educating them about the enormity and complexity of the Internet, including initiation of such users to the idea that freedom of expression online often crosses over to hateful speech and harassment. This would act as an initiation process to understand the working of the Internet and the prevalence of hateful and harassing content on its numerous speech platforms, so that first-time users are not discouraged from using the Internet merely due to the presence of negative content. An interesting suggestion for the platforms was to work towards a mechanism that is more offender centric, and facilitates the tracking of repeat offenders along with providing tools of blocking for users.

This roundtable served in exploring the many layers of hateful and harassing speech that runs across roles and responsibilities of various stakeholder groups and concerns that are deeply entrenched in our societal outlook. The increasing frequency and amount of such content on the Internet is an indication of the urgent need to collaborate and develop a framework for limiting such speech, while balancing the Fundamental Right to freedom of expression.

4.2 Second roundtable (New Delhi; September 6, 2016)

On September 6, 2016, SFLC.in organized the second in a series of consultations at the India International Centre, New Delhi on the various facets of hateful and harassing speech online. Both the roundtables had representation from industry, civil society, and media actors, and were instrumental in achieving a nuanced and deeper understanding of what constitutes harmful speech on an online forum, the challenges that intermediaries face in moderating such content, and the roles and responsibilities of LEAs. A set of draft best practices that could be adopted by the intermediaries as a self regulatory measure were proposed by SFLC.in and were discussed at length over the course of the two roundtable discussions. Please note that both roundtables were held under the Chatham house rules to facilitate an open exchange of ideas, and no attributions will be made as to the sources of viewpoints discussed below.

Over the course of the second discussion, it was highlighted that it is indeed difficult to capture the contours of online harassment and hate speech in definite terms, as the line between legitimate and abusive uses of the freedom of expression is subjective. In addition, social media is often the chosen platform for powerful players to further their propaganda, and these positions of power are at times used to popularize certain kinds of opinions at the expense of others. With regards to free speech and expression in the online space, anonymity poses a unique conundrum, where on one hand it facilitates free and open discourse amongst vulnerable groups and minorities, while on the other, it is used as a mask by the perpetrators of harassing and abusive speech.

Although the most widely used social media platforms have policies that strongly condemn and restrict the use of their networks for abusive and harassing speech, it was discussed that these terms of service and community standards prove to be problematic for both the user and the intermediary owing to the lack of objective criteria to determine the extent of restricted content on particular platforms. This results in a situation where the user is unable to determine if their opinions would be violative of the set standards, and the intermediaries are caught between censoring too much, or not censoring enough. From a user perspective, it was suggested that due to the large volume of information available about users on certain types of platforms, intermediaries should probably develop mechanisms wherein they can enhance protection for information they retain, especially about vulnerable groups and minorities. To improve transparency on the part of intermediaries, it was recommended that a comprehensive explanation about the reasons for removal of particular content should be provided. For example, if the filtration is done through algorithms, the ‘phrases’ or words in the text, or graphics in the image that were flagged by the algorithm should be mentioned to better understand the working of community standards and content moderation policies.

It was also pointed out that the policies and standards developed by the platforms are not set in stone, and that the tools for customizing various platforms according to user needs evolve with public consultations with various groups and organizations. However, a lack of user awareness and know-how in the usage of the existing tools for blocking, muting, and reporting was unanimously acknowledged by all stakeholders present, and hence it was mentioned that various platforms are conducting campaigns on these fronts are ongoing especially amongst vulnerable groups and rural communities. A suggestion for automated filtering of entire phrases that could constitute as harassing and hateful was dismissed as being obstructive of legitimate free speech as well. To refine their practices with the expanding base of global users, certain intermediaries are using language experts to ensure that harassing and abusive content is removed from their platforms. Therefore, efforts are underway by the platforms to develop tools and better the mechanisms for detecting abusive content, as well as provide filters and tools for users to employ.

V. State responses

While the most debated responses to hate speech – offline and online – have primarily centered around law, a strictly legal approach has its risks and limitations. *First*, as any relevant regulation in this regard would inevitably have to dabble in restricting the freedom of speech in larger public interest, there is the ever-present danger of placing collateral restrictions on legitimate expression. This fear is all the more valid when considering the recognition accorded by some jurisdictions to the individual's right to "offend, shock or disturb others", which further complicates the difficulties inherent in clearly outlining the parameters of hate speech.[28] *Second*, there is the fact that law is in fact enforcing the mores of the dominant group that controls the content of the law, despite its projection as simply enforcing the given and natural norms of a decent society.[29] For instance, hate speech law in Apartheid South Africa was used to criminalize criticism of white domination, which illustrates the potential political abuse of hate speech limitations prescribed by law.[30] *Third*, a purely legal lens can miss out on how societies evolve through contestation and disagreement. Although hate and intolerance are offensive and low expressions of dissent, they can also be thought of as windows into deeply-rooted tensions and inequalities, which themselves do need addressing beyond pure speech issues, and beyond the online dimension.[31] In light of the above, while this chapter will examine India's state responses to hateful and harassing speech online, it becomes necessary to

[28] Supra. 5

[29] Supra. 6, p. 15

[30] ibid.

[31] Ibid.

place particular emphasis on responses other than state-initiated legal measures.

India’s state responses to online harassment have been centered almost exclusively around the law. A comprehensive picture of the laws relevant in this regard may be gleaned by examining them under two broad heads, namely laws that protect and promote free speech, and other laws that prescribe specific civil and criminal remedies.

5.1 Protection and promotion of free speech

Article 19(1)(a) of the **Constitution of India** states, *“All citizens shall have the right…to freedom of speech and expression”.*[32]

Under Article 19(1)(a), the Constitution of India guarantees to all its citizens the Fundamental Right to Freedom of Speech and Expression, which according to **Article 19(2)** can be reasonably restricted by law only in the interests of the sovereignty and integrity of India, security of the State, friendly relations with foreign States, public order, decency or morality, or in relation to contempt of court, defamation or incitement to an offense.[33]

Several judgments of the Supreme Court of India have referred to the importance of this Fundamental Right – both from the point of view of the liberty of the individual and from the point of view of India’s democratic form of Government. For example, in the early case of *Romesh Thappar v. State of Madras*[34] the Supreme Court stated that freedom of speech lay at the foundation of all democratic organizations. In *Sakal Papers (P) Ltd. and Ors. V. Union of India*[35], a Constitution Bench of the Supreme Court said freedom of speech and expression of opinion is of paramount importance under a democratic Constitution, which envisages changes in the composition of legislatures and Governments, and must be preserved. In a separate concurring judgment in *Bennett Coleman & Co. and Ors. v. Union of India and Ors.*[36], the Supreme Court said that the freedom of speech and of the press is the Ark of the Covenant of Democracy because public criticism is essential to the working of its institutions.

Further, India is a signatory to the Universal Declaration of Human Rights (UDHR) as well as the International Covenant on Civil and Political Rights (ICCPR), both of which contain provisions that recognize the individual’s right to free speech and expression.

Notably, Article 19 of the UDHR states:

[32] Article 19(1)(a), Constitution of India, 1949

[33] Ibid.

[34] 1950 (1) SCR 594

[35] 1962 (3) SCR 842

[36] 1973 (2) SCR 757

> *Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.*[37]

Similarly, Article 19 of the ICCPR states:

> 1. *Everyone shall have the right to hold opinions without interference.*
>
> 2. *Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice.*
>
> 3. *The exercise of the rights provided for in paragraph 2 of this article carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary:*
>
>   *a.* *For respect of the rights or reputations of others;*
>
>   b. *For the protection of national security or of public order (ordre public), or of public health or morals.*[38]

As noted by David Kaye, United Nations (UN) Special Rapporteur on the Promotion and Protection of the Right to Freedom of Expression and Opinion, opinion and expression are closely related to one another, as restrictions on the right to receive information and ideas may interfere with the ability to hold opinions, and interference with the holding of opinions necessarily restricts the expression of them.[39] However, it may be worth noting that human rights law draws a conceptual distinction between the two. During the negotiations on the drafting of the ICCPR, the freedom to form an opinion and to develop this by way of reasoning was held to be absolute and, in contrast to freedom of expression, not allowed to be restricted by law or other power.[40] The right to freedom of expression under Article19 of the ICCPR expands upon the UDHR's already broad guarantee, protecting the freedom to seek, receive and impart information and ideas of all kinds, regardless of

[37] Article 19, Universal Declaration of Human Rights, available at: http://www.un.org/en/universal-declaration-human-rights, last accessed on November 14, 2016

[38] Article 19, International Covenant on Civil and Political Rights, available at: http://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx, last accessed on November 14, 2016

[39] United Nations, Human Rights Council, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye*, A/HRC/17/27, p. 8, available at: http://www.ohchr.org/EN/HRBodies/HRC/RegularSessions/Session29/Documents/A.HRC.29.32_AEV.doc, last accessed on January 4, 2016

[40] Manfred Nowak, *UN Covenant on Civil and Political Rights: CCPR Commentary* (1993), p. 441.

frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice. A significant accumulation of jurisprudence, special procedure reporting, and resolutions within the UN and regional human rights systems underscores that the freedom of expression is essential for the enjoyment of other human rights and freedoms, and constitutes a fundamental pillar for building a democratic society and strengthening democracy.[41]

5.2 Other legal remedies

Until as recently as March 2015, most variants of online harassment as discussed over the previous chapters would have constituted offenses under **Section 66A** of the **Information Technology Act, 2000**. However, the Supreme Court of India in March 2015 struck down Section 66A as unconstitutional, finding that it was violative of free speech. When in force, Section 66A read:

> *66A – Punishment for sending offensive messages through communication service, etc. – Any person who sends, by means of a computer resource or a communication device;*
>
> *a. any information that is grossly offensive or has menacing character; or*
>
> *b. any information which he knows to be false, but for the purpose of causing annoyance, inconvenience, danger, obstruction, insult, injury, criminal intimidation, enmity, hatred or ill will, persistently by making use of such computer resource or a communication device; or*
>
> *c. any electronic mail or electronic mail message for the purpose of causing annoyance or inconvenience or to deceive or to mislead the addressee or recipient about the origin of such messages,*
>
> *shall be punishable with imprisonment for a term which may extend to three years and with fine.*
>
> *Explanation. -For the purpose of this section, terms "electronic mail" and "electronic mail message" means a message or information created or transmitted or received on a computer, computer system, computer resource or communication device including attachments in text, image, audio, video and any other electronic record, which may be transmitted with the message.*

Visibly broad and ambiguous terms such as “grossly offensive”, “menacing character”, “annoyance” and “inconvenience” were not defined under any Indian legislation including the IT

[41] Supra. 39, pp. 8 – 9

Act, meaning they remained highly subjective terms of every-day parlance, whose applicability to alleged infringements of Section 66A varied greatly from person to person. Coupled with the fact that no warrant was required to make an arrest under Section 66A as it was a cognizable offence, the Section lent itself to wanton abuse. Soon after it was introduced into the IT Act by way of a 2012 Amendment, a number of instances were reported, where seemingly innocent citizens were charged with offenses under this provision:

- *April 2012:* Jadavpur university professor Ambikesh Mahapatra and his neighbor Subrata Sengupta were arrested for allegedly circulating a cartoon that lampooned West Bengal chief minister Mamta Banerjee. The cartoon, which was widely circulated on the Internet, was based on a scene in a film in which a boy was duped by two criminals into believing that they caused someone to vanish. In the cartoon, the vanishing man was a reference to former railways minister Dinesh Trivedi, who was forced out of office by Mamta Banerjee.[42]

  Both Mahapatra and Gupta were booked on charges of defamation, outraging the modesty of a woman and hacking. [Sections 500, 509 of the Indian Penal Code, 1973 (IPC) and Sections 66A and 66B of the IT Act][43]

- *May 2012:* Free speech campaigner and cartoonist Aseem Trivedi was arrested in Mumbai for displaying cartoons on his website and Facebook page that mocked parliament and corruption in high places. The caricatures were shared on other social media. Trivedi's cartoons purportedly depicted the parliament as a giant commode and showed the national emblem with wolves instead of lions.

  He was charged with sedition under Section 124 A of the Indian Penal Code, the Prevention of Insults to National Honour Act and Section 66A of the IT Act. [44]

- *November 2012:* Two young girls from Palghar, Mumbai, were arrested when one of them posted a question on her Facebook page questioning why the city was shut down for Shiv Sena leader Bal Thackeray's funeral. One of them commented that the shutdown was out of fear, not respect. The second girl, her friend, was arrested for liking the post.

  They were arrested for "hurting religious sentiments" under Section 295(a) of the IPC and

[42] *Facebook trouble: 10 cases of arrests under Sec 66A of IT Act*, Hindustan Times, 24th March 2015, available at: http://www.hindustantimes.com/india/facebook-trouble-10-cases-of-arrests-under-sec-66a-of-it-act/story-4xKp9EJjR6YoyrC2rUUMDN.html, last accessed on January 7, 2016

[43] Jadavpur University Professor Arrested in Kolkata, available at: http://sflc.in/chilling-effects/police-action-66a/jadavpur-university-professor-arrested-in-kolkata/, last accessed on January 7, 2016

[44] Supra. 42

Section 66(a) of the IT Act. All charges were later quashed by a court.[45]

- *August 2013:* Poet and writer Kanwal Bharti was arrested by police for posting a message on Facebook that criticized the Uttar Pradesh government for suspending IAS officer Durga Shakti Nagpal, who had cracked down on the sand mafia. Bharti's post on Facebook questioned why Nagpal had been suspended for ordering the demolition of a wall intended to be part of a mosque while no officer in Rampur was dismissed when an old *madrassa* was pulled down.[46]

- *May 2014:* Ship-building professional Devu Chodankar was arrested for posting a comment against Prime Minister Narendra Modi on Facebook. Police described Chodankar's comment as part of a "larger game plan to promote communal and social disharmony in the state"

  Police filed an FIR against him under Sections 153(A) and 295(A) of the IPC and section 125 of the Representation of the People's Act, and Section 66A of the IT Act.[47]

- *March 2015:* A teenager of Class XI was arrested in Rampur and sent to jail for allegedly posting on Facebook "objectionable" comments attributed to Uttar Pradesh Minister Azam Khan, kicking off a fresh controversy over the booking of people under Section 66A. The youth was later released on bail and the Supreme Court sought explanation from UP Police on the circumstances leading to the arrest.[48]

All the widely publicized arrests came to a head when the constitutional validity of Section 66A was challenged before the Supreme Court of India by 10 separate petitions, starting with *Shreya Singhal and Ors. v. Union of India*[49] in 2012. Though some of these petitions also raised challenges to other provisions of the IT Act, the primary focus was on Section 66A, whose Constitutional standing was questioned on the following broad grounds:[50]

i. It violated the Constitutionally guaranteed Fundamental Right to Freedom of Speech and Expression under Article 19(1)(a), because:

   o the restrictions it imposed on the right guaranteed under Article 19(1)(a) are beyond

[45] Supra. 43

[46] Ibid.

[47] Ibid.

[48] Ibid.

[49] AIR 2015 SC 1523

[50] Compiled from the Supreme Court's summation of the petitioner's arguments in its judgment, available at: http://supremecourtofindia.nic.in/FileServer/2015-03-24_1427183283.pdf. A more detailed reporting of the 9 petitions clubbed with *Shreya Singhal* is available at: http://sflc.in/information-technology-act-and-rules-time-to-change/

the scope of permissible restrictions enumerated under Article 19(2)

- it suffers from vagueness as terms such as “annoying”, “inconvenience” and “grossly offensive” are not defined under any Indian legislation, the result being that innocent persons are roped in as well as those who are not
- its enforcement would be an insidious form of censorship, which impairs a core value contained in Article 19(1)(a)
- it has a chilling effect of freedom of speech and expression
- it deprives viewers of the multiplicity of views that could be accessed over the Internet

ii. It violated the Constitutionally guaranteed Fundamental Rights to Equality before Law, and Life and Personal Liberty guaranteed under Articles 14 and 21 of the Constitution respectively, because it discriminated between those using the Internet and those using other means of communication to commit alleged infringements, while no such *intelligible differentia* actually exists

In a historic verdict delivered in March 2015, the Supreme Court struck down the provision as unconstitutional, as it was found to violate the Fundamental Right to Freedom of Speech and Expression guaranteed under Article 19(1)(a) of the Constitution, and was not saved by the reasonable restrictions permissible under Article 19(2).

In its judgment, the Supreme Court observed that liberty of free speech and expression are cardinal values of paramount significance to the constitutional process in democracies. Mere discussion or even advocacy of a cause, howsoever unpopular, are at the heart of Article 19(1)(a) of the Constitution, and it is only when such discussion or advocacy reaches the level of incitement that Article 19(2) kicks in. The Court, unconvinced by the Government’s assurance that Section 66A would only be used in a responsible manner, held that the Section not only failed the ‘clear and present danger’ test, but also bore no proximate relation to any of the subject matters enumerated under Article 19(2), especially to public order. Moreover, the Court found every expression used in Section 66A to be nebulous and imprecise, and held that the global reach of the Internet can neither restrict the content of Article 19(1)(a), nor justify its denial. As a result, Section 66A was held to be vague, over-broad, violative of Article 19(1)(a), and not saved by Article 19(2). The Section was accordingly struck down as unconstitutional, marking the end of an era, albeit a short one, in India's regulation of online content.

The repeal of Section 66A might have de-clawed the Indian legal machinery to a certain extent when it comes to combating online harassment, but it was clearly a necessary sacrifice. Even with Section 66A gone, several other provisions remain across Indian statutes, penalizing hateful and harassing speech of varying kinds. Most of these provisions co-existed with Section 66A, and several of the penal provisions were often invoked over and above charges under Section 66A. In fact, one of the contentions raised before the Supreme Court by the petitioners in *Shreya Singhal v. Union of India* was that most offenses that Section 66A sought to punish were already covered under existing legislations such as the **Indian Penal Code, 1860**, which are medium-neutral.

Broader, more socially targeted instances of harmful speech – as seen during the violence surrounding the beef bans of 2015 – will attract sanctions under provisions of the IPC such as:

- **Section 153A (1)** [promoting enmity between different groups on grounds of religion, race, place of birth, residence, language, etc. and doing acts prejudicial to maintenance of harmony; imprisonment up to 3 years or fine or both] – Applies to:
    - Words (spoken/written), signs, visible representations promoting disharmony or feelings of enmity/hatred/ill-will between groups on any grounds, including but not limited to religion, race, place of birth, residence, language, caste or community
    - Acts likely to cause disharmony among religious/racial/language/regional groups, and acts likely to disturb public tranquility
- **Section 153B (1)** [imputations, assertions prejudicial to national-integration; imprisonment up to 3 years or fine or both] – Applies to:
    - Words (spoken/written), signs, visible representations that impute inability to bear allegiance to the Constitution or uphold the sovereignty and integrity of India by virtue of membership in religious/racial/language/regional groups; advocate the denial of citizens' rights to members of religious/racial/language/regional groups;
- **Section 295A** [deliberate and malicious acts, intended to outrage religious feelings by insulting its religion or religious beliefs; imprisonment up to 3 years or fine or both] – Applies to:
    - Words (spoken/written), signs, visible representations with deliberate and malicious intention that insult the religious feelings of any class of citizens
- **Section 505** [statements conducting to public mischief; imprisonment up to 3 years or fine

or both] – Applies to:

    - Statements, rumors, reports intended or likely to cause public alarm, whereby any person is induced to commit an offense against the State or against public tranquility; statements, rumors, reports intended or likely to incite classes or communities to commit offenses against one another
    - Statements, reports containing rumors or alarming news intended or likely to cause feelings of enmity/hatred/ill-will among religious/language/racial/regional groups on any grounds including religion, race, place of birth, residence, language, caste or community

Individually targeted instances of harmful speech, including online harassment on the other hand, will attract sanctions under:

- **Section 354D** [stalking] – Applies to:
    - Monitor of a woman's use of the Internet, email or any other form of electronic communication
- **Section 503** [criminal intimidation; imprisonment up to 7 years or fine or both] – Applies to:
    - Threats of injury to one's person/reputation/property or to the person/reputation of another in whom one is interested, with intent to cause alarm, force one to perform a legally non-obligatory act or omit a legally entitled act
- **Section 504** [intentional insult with intent to provoke breach of the peace; imprisonment up to 2 years or fine or both] – Applies to:
    - Insults intended or known to be likely to provoke one to breach public peace or commit an offense
- **Section 507** [criminal intimidation by an anonymous communication; penalty for criminal intimidation + imprisonment for up to 2 years or fine or both] – Applies to:
    - Criminal intimidation with added precautions to conceal the name or abode of the perpetrator

VI. Non-state responses

In view of the risks and limitations involved in over-reliance on legal-centric responses to harassing speech online, it becomes important that there be an equal or greater number of non-legal-centric, non-state-initiated efforts that address the issue. Non-state responses in this context would most notably include preventive and remedial frameworks instituted by platforms where users run the risk of falling victim to online abuse, and dedicated campaigns that bring together experts and other relevant stakeholders to offer support, conduct effective dialogue and propose/initiate measures that contribute to limiting instances of online abuse. The following pages will examine these twin components of non-state responses to online harassment in greater detail.

6.1 Content reporting mechanisms

Many social media platforms have experienced heat from the public for, at times not promptly removing abusive content, or on the contrary for censorship of unpopular content. Most social networking websites have an established set of standards in their Terms of Service and related policy documents that enumerate the kind of content not permissible on their forum, and they usually provide express guidelines for users to report content they think is violating the standards. Below are brief overviews of the nature and enforcement of such policies by a few popular social media platforms.

6.1.1 Twitter

Twitter was founded in 2006 as a micro-blogging platform that allowed users to express themselves and share content of their choice in under 140 characters. The service rapidly gained worldwide popularity, with more than 100 million users posting 340 million tweets a day in 2012. The service also handled 1.6 billion search queries per day. In 2013, Twitter was one of the ten most-visited websites and has been described as "the SMS of the Internet". As of May 2015, Twitter had more than 500 million users, out of which more than 332 million were active.

Over a decade of its existence, Twitter has gained both popularity as a crucial enabler of online free speech and a notoriety for lending itself to wanton abuse of this enablement. For instance, a 2014 study conducted by Kick It Out – an organization that works for equality and inclusion in the game of football – revealed that around 134,400 football-related discriminatory posts were made on social media between August 2014 and March 2015, of which 88% came from Twitter.[51] Facebook on the other hand hosted 8% of discriminatory posts, with other online fora and blogs making up the remaining 4%. While these figures may not indicate Twitter's contribution to the general state of online abuse, it does show that the platform, by virtue of its encouragement of spontaneous publication of bite-sized content, lends itself to abuse more than other social media platforms.

[51] *Kick It Out unveils findings of research into football-related hate crime on social media*, April 16, 2015, available at: http://www.kickitout.org/news/kick-it-out-unveils-findings-of-research-into-football-related-hate-crime-on-social-media/#.VvjWkTbwwb1, last accessed on March 28, 2016

6.1.1.1 Relevant policies

Despite being launched to the public in 2006, it wasn't until 2009, when Twitter had amassed a user-base of around 5 million, that its Terms of Service and "Twitter Rules" were introduced.[52]

Version 1 of the Terms of Service[53], a substantially more concise document than it is today, contained three clauses under the head of "Basic Terms" that collectively mandated users to refrain from abusive behavior and illegal activities, held international users responsible for compliance with local laws regarding permissibility of content, and assigned sole responsibility for content posted on Twitter to the respective users themselves. The document also included a clause under the head "General Conditions", which reserved Twitter the right (not an obligation) to remove content or accounts containing content that were determined at its sole discretion to be unlawful, offensive, threatening, libelous, defamatory, obscene or otherwise objectionable or violative of any party's intellectual property or the Terms of Service.

The accompanying Twitter Rules were also rather brief (568 words), and reflected Twitter's policy of forbearance stating, "we do not actively monitor user's content and will not censor user content, except in limited circumstances".[54] The Rules also laid out 10 separate heads of "content boundaries", of which about 4 dealt with hate, intolerance and harassment. As per the content boundaries, impersonation intended to "mislead, confuse or deceive", direct and specific threats of violence, use of Twitter for "unlawful purposes or for promotion of illegal activities", and creation of serial accounts for "disruptive or abusive purposes" were prohibited.

In December 2015, Twitter introduced overarching changes to the Twitter Rules, adding 178 new words to the document.[55] While none of these changes constituted broader policy changes per se, they nevertheless served to coalesce effective bans on hate speech, incitement of harassment etc. under one umbrella document.

As of today, the Twitter Rules contain two additional heads as compared to Version 1 of the document introduced in 2009, namely "abusive behavior" and "spam". The "abusive behavior" head is of particular interest to limiting hate, intolerance and harassment online, as the following activities now feature as grounds that may attract temporary locks and/or permanent suspension of

[52] Sarah Jeong, *The History of Twitter's Rules*, 14th January 2016, available at: http://motherboard.vice.com/read/the-history-of-twitters-rules, last accessed on March 2, 2016

[53] Twitter Terms of Service, Version 1, 2009, available at: https://twitter.com/tos/previous/version_1?lang=en, last accessed on March 2, 2016

[54] The Twitter Rules, Version 1, 14th January 2009, available at: https://web.archive.org/web/20090118211301/http://twitter.zendesk.com/forums/26257/entries/18311, last accessed on March 22, 2016

[55] Supra. 55

user accounts that contravene the Twitter Rules:

- Threats of violence or promotion of violence, including threats or promotion of terrorism
- Incitement of or engagement in targeted abuse or harassment of others, including by:
    - Creation of accounts whose primary purpose is to harass or send abusive messages
    - Engaging in one-sided behavior including threats
    - Sending harassing messages to an account from multiple accounts
- Promotion of violence against or direct attacks or threats against other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease
- Creation of multiple accounts in order to evade temporary or permanent suspension of particular accounts
- Publishing or posing other people's private and confidential information, such as credit card numbers, street address, or Social Security/National Identity numbers, without their express authorization and permission
- Posting intimate photos or videos that were taken or distributed without the subject's consent
- Impersonation that is intended to or does mislead, confuse, or deceive others

6.1.1.2 Enforcement of policies

Version 1 of both the Terms of Service and Twitter Rules were more or less spartan. They were indicative of Twitter's initial projected image as an open platform that enabled users to share bite-sized content with minimal oversight and little to no censorship. Protocols governing its use were relatively broad and non-specific, and interventions from Twitter were sought to be kept at a bare minimum – all policies that earned it a reputation for being free-speech friendly. Over the subsequent years however, usage of the platform saw a significant upswing and Twitter went from a novel concept used by a few to a pervasive social media platform that is deeply integrated into Internet use even outside of Twitter.

With such exploding adoption of Twitter as a speech platform came a slew of high-profile conflicts and controversies that forced the company to rethink its content oversight policies so as to better shield its millions of users from the increasing risk of harm. Notable amongst the controversies that prompted changes were a flood of rape threats reported by British feminists in 2013[56], Zelda Williams' abrupt exit from Twitter in 2014 over insensitive and harassing messages following her father Robin Williams' passing[57], and the infamous #Gamergate scandal from the same year when several female executives in the gaming industry were subjected to relentless harassment online.[58]

Building pressure from various quarters saw Twitter make a few feature additions and under-the-hood changes to the Twitter Rules over the years, mostly geared towards addressing policy shortcomings brought to light by specific controversies. The introduction of the "Report abuse" button, ability to export and share "blocked lists", and the introduction of "targeted abuse" as a prohibited activity under the Twitter Rules are all great examples of the heightened anti-hate protocols effectuated in response to particular controversies.

6.1.2 Facebook

With roughly 1.5 billion active user-accounts, Facebook is the single largest social networking platform in the world, and bears a heavier-than-usual burden in terms of safeguarding users from harm. Britain's largest police force, the Metropolitan Police, reported having received 1,207 crime reports which mentioned Facebook in 2014, up from 935 in 2013 and 997 in 2012.[59] Similarly, the Police in Queensland, Australia had over 5,000 cases last year that, in some way, involved Facebook - a 50 percent increase from the previous year.[60]

6.1.2.1 Relevant policies

Facebook's policies against hate, intolerance and harassment are distributed over three sets of documents, each serving distinct underlying functions. These are: Statement of Rights and Responsibilities; Data Policy; and Community Standards. Below are brief overviews of included provisions that deal specifically with hate, intolerance and harassment.

- **Statement of Rights and Responsibilities (SRR):** This effectively constitutes the Terms of

[56] *Woman who campaigned for Jane Austen banknote receives Twitter death threats*, The Telegraph UK, 28th July 2013, available at: http://www.telegraph.co.uk/technology/10207231/Woman-who-campaigned-for-Jane-Austen-bank-note-receives-Twitter-death-threats.html, last accessed on March 4, 2016

[57] Cassandra Khaw, *Zelda Williams leaves Twitter on account of social media abuse*, The Verge, 13th August 2014, available at: http://www.theverge.com/2014/8/13/5997743/zelda-williams-leaves-twitter, last accessed on March 4, 2016

[58] Jay Hathaway, *What is Gamergate and Why? An Explainer for Non-Geeks*, Gawker, 10th October 2014, available at: http://gawker.com/what-is-gamergate-and-why-an-explainer-for-non-geeks-1642909080, last accessed on March 4, 2016

[59] *Police facing rising tide of social media crimes*, The Telegraph UK, June 5, 2015, available at: http://www.telegraph.co.uk/news/uknews/crime/11653092/Police-facing-rising-tide-of-social-media-crimes.html, last accessed on March 24, 2016

[60] *Shocking statistics on Facebook related crimes*, November 6, 2012, available at: http://facecrooks.com/Internet-Safety-Privacy/Shocking-Statistics-on-Facebook-Related-Crimes.html/, last accessed on March 24, 2016

Service that governs Facebook's relationship with users and others who interact with the platform. Every individual that uses or accesses Facebook and its services agrees to the SRR by default, and violation of its letter or spirit may result in the stoppage of Facebook's services to the offending party. The SRR contains several user mandates that lay the groundwork for safeguarding Facebook users from instances of abuse, including the following:[61]

- Do not procure login information or access someone else's account
- Do not bully, intimidate or harass any user
- Do not post content that is hate speech, threatening, or pornographic; incites violence; or contains nudity or graphic or gratuitous violence
- Do not use Facebook to do anything unlawful, misleading, malicious, or discriminatory
- Do not post content or take any action on Facebook that infringes or violates someone else's rights or otherwise violates the law
- Do not post anyone's identification documents or sensitive financial information on Facebook

The SRR also contains a rather controversial rule that users must provide their authentic names and information while signing up for Facebook. In that spirit, users are also prohibited from providing any false personal information on Facebook, creating an account for anyone other than themselves without permission, or creating more than one personal account. The "real name policy" has drawn significant criticism from certain segments of Facebook's user-base as it exposes vulnerable users to harm by allowing easy identification. However, Facebook has continued to maintain that this measure is essential to ensure accountability and users' safety.[62]

- **Data Policy:** This document effectively serves as Facebook's Privacy Policy, and describes what information it collects from its users and how such information is used and shared. Apart from detailing the specific kinds of information Facebook collects from users, the broad uses to which such information is put and the parameters under which the information

[61] Facebook, *Statement of Rights and Responsibilities*, available at: https://www.facebook.com/legal/terms, last accessed on March 24, 2016

[62] Russell Brandom, *Facebook is changing the way it enforces the real name policy*, The Verge, 15th December 2015, available at: http://www.theverge.com/2015/12/15/10215936/facebook-real-name-policy-changes-appeal-process, last accessed on March 24, 2016

is shared with third-parties, the Data Policy, under the head “How do we respond to legal requests or prevent harm?”, provides that Facebook may access, preserve and share information when it is considered necessary to: detect, prevent and address fraud and other illegal activity; protect itself, its users and others, including as part of investigations; or prevent death or imminent bodily harm. Facebook also reserves the right to retain information from accounts disabled for violations of its terms for at least a year to prevent repeat abuse or other violations of its terms.[63]

- **Community Standards:** Loosely comparable to codes of conduct, Community Standards are essentially a detailed set of guidelines designed to help users understand what content is and isn’t considered acceptable on Facebook, and are perhaps the most illustrative of Facebook’s own determinations of the kinds of content that need to be filtered from public view. The guidelines are presented under four sub-heads that address personal and public safety, respectful behavior, security of user accounts and personal information, and protection of Intellectual Property, of which the first three are relevant to limiting hate, intolerance and harassment. The following kinds of abusive content and activities are proscribed as per Facebook’s Community Standards, and will be removed from public access upon detection:
  - Credible threats of physical harm to individuals, including specific threats of theft, vandalism and other financial harm; factors such as physical location and public visibility contribute to determining credibility of threats, and all threats may be presumed credible in violent/unstable areas
  - Presence (user accounts, pages etc.) of organizations involved in terrorist/organized criminal activity (dangerous organizations)
  - Expression of support for dangerous organizations and groups, including by supporting or praising their leaders or condoning their violent activities
  - Purposeful targeting of private individuals (people who have not gained news attention or interest of the public) to degrade or shame them, including by creating pages, posting altered images, posting photos/videos of physical bullying, sharing personal information to blackmail and harass, repeatedly sending friend requests or messages

[63] Facebook Data Policy, available at: https://www.facebook.com/about/privacy, last accessed on March 24, 2016

- Facilitation of any manner of criminal activity, celebration of crimes committed
- Threats or promotions of sexual violence/exploitation including sexual exploitation of minors and sexual assaults, photos/videos depicting sexual violence, images shared in revenge without consent; Facebook defines sexual exploitation as solicitation of sexual material, any sexual content involving minors, threats to share intimate images and offers of sexual services
- Hate speech i.e. content that directly attacks people based on their race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender or gender identity, or serious disabilities or diseases
- Presence of organizations and people dedicated to promoting hatred against the groups mentioned above

6.1.2.2 Enforcement of policies

Facebook offers its users multiple means of dealing with content that they find offensive, inappropriate, dangerous or otherwise undesirable. Offending users may be "un-friended", leaving them unable to chat with or post messages on their victim's timeline. Such users may also be blocked by their victims, which would render them unable to add their victims as friends or view content on their timelines. In 2014, Facebook launched systems to allow people to directly engage with one another to better resolve their issues beyond simply blocking or un-friending another user. Of particular note, is the "social reporting tool" that allows people to reach out to other users or trusted friends to help resolve conflicts or start a dialogue about a piece of content.

Most important amongst the available means to limit hate, intolerance and harassment however are the "report abuse" buttons found at several points across the Facebook platform, allowing content to be flagged as possibly violative of Facebook policies. Facebook does monitor its networks suo moto and remove content found in violation of its policies, but a vast majority of content removals are nevertheless the result of abuse reports made by individual users. Four dedicated "User Operations" teams – two of which are located in the United States, one in Ireland and one in India – handle the millions of abuse reports received each day by Facebook in 24 different languages. These teams are situated across the world in such a way as to ensure that at least one Facebook team is handling reports at all times. Each User Operations team is separated into four sub-teams based on the kind of reports each one handles - the Safety team, the Hate and Harassment team, the Access team, and the Abusive Content team. When a person reports a piece of content, depending on the reason for their report, it will go to one of these teams. For instance, content reported to contain graphic violence, will be reviewed and assessed by the Safety Team. Users are able to keep track of the reports they make via the "Support Dashboard" feature available on their Facebook home page.[64]

If one of the User Operations teams determines that a reported piece of content violates Facebook policies, the content will be removed and the user who posted it will be notified. In addition to this, Facebook may also revoke a user's ability to share particular types of content or use certain features, disable their account, or refer issues to law enforcement. Facebook also employs special teams to handle user appeals against actions taken.[65]

Though Facebook does have an elaborate reporting mechanism in place when it comes to instances of online harassment, there have also been numerous criticisms about the transparency of this process and its algorithmic undertone. Though Facebook maintains that this process is not effected by the amount of reports received against particular content, many incidents have brought this stand to question. This 'report abuse' button has been used notoriously by many to censor political speech and shut down pages with unpopular views.

It was reported that one of Vietnam's independent news sites, Khmer Krom News' page was bombarded with planned 'report abuse' requests from a pro- Government site as a tactic of shutting down the page that published critical views regarding the Government. Moreover, the same strategy was used to close accounts of 44 activists and journalists in the country. It is still not clear what community standard these people or pages had violated.[66] In another incident on Instagram (also owned by Facebook), a woman's photograph with a menstrual stain was deleted twice for contravening the guidelines prescribed for using Instagram.[67] Through these repeated instances, questions have arisen on the effectiveness of this tool as a means for reporting harmful content or a weapon for private censorship. In another incident (also narrated under chapter III of this report), two Indian women named Preetha G Nair and Inji Pennu suffered from multiple threats, bullying and harassment by online trolls after speaking about a politician on Facebook. In turn, Facebook suspended their accounts with a automated message saying it had been reported as against their

[64] Facebook, *What happens after you click "Report"?*, 9th June 2012, available at: https://www.facebook.com/notes/facebook-safety/what-happens-after-you-click-report/432670926753695, last accessed on March 24, 2016

[65] Supra. 64

[66] Russell Brandom. *Facebook's Report Abuse button has become a tool for global oppression*, The Verge, September 2, 2014, available at: http://www.theverge.com/2014/9/2/6083647/facebook-s-report-abuse-button-has-become-a-tool-of-global-oppression, last accessed on March 5, 2016

[67] Radhika Sanghani , *Instagram deletes woman's period photos-but her response is amazing*, The Telegraph, March 30, 2015, available at: http://www.telegraph.co.uk/women/life/instagram-deletes-womans-period-photos-but-her-response-is-amazing/, last accessed on March 5, 2016

community standards, or the account did not use the real name (as is Facebook's policy).[68] It is not difficult for a troll to create a new account or page if their previous accounts have been suspended from Facebook. As per Facebook's policy, multiple violations of community standards will be determined on the basis of severity of act as well as the history of the person. Nevertheless, with incidents like Inji Pennu, where the complainant's account is suspended and the trolls continue to appear, it is uncertain how this social media giant tackles their report abuse requests and this whack-a-mole situation. Transparency and discussions on these issues are highly desired.

6.1.3 Reddit

As a popular discussion board known for its commitment to free speech, Reddit prohibits content that is illegal, involuntarily pornographic, incites violence, threatens, harasses or bullies, or does not tag nudity, profanity or pornography as NSFW (Not Safe For Work).[69] Users are encouraged to use the "report" button built into comments and links to bring to the moderators' attention content identified as spam or otherwise violative of the Reddit rules. This report may be accompanied by a written explanation (under 100 characters) as to why the content was thought to be violative. For longer explanations of content reports, users must leave replies or send messages to the moderators[70] These moderators are chosen by moderators of other sub-reddits on the basis of various factors, including their frequency and performance on the specific thread, the age of their account on Reddit, the effort they make in their post/comments on Reddit, and other factors.[71] Recently, Reddit was under a lot of pressure to edit their policies due to the events with certain sub-reddits that promoted hate speech and bullying.[72] But the moderators of two of the most famous Reddit threads revealed that they follow two kinds of approaches for moderating such content on their forum.[73] One, is the complete banning of the community in question. This was used against the sub-reddit r/fatpeoplehate. But as an after effect, the members of that community make it their mission to bombard other unrelated communities with their opinions, thereby causing an influx of

[68] Sahar Habib Ghazi, *"We will Choke You": How Indian women face fatal threats on Facebook while Trolls roam free*, August 6, 2015, Global Voices Advocacy; available at: https://advox.globalvoices.org/2015/08/06/we-will-choke-you-how-indian-women-face-fatal-threats-on-facebook-while-trolls-roam-free/; https://globalvoices.org/2015/08/02/indians-blast-facebook-over-broken-community-standards/, last accessed on August 17, 2016

[69] Reddit Content Policy, available at: https://www.reddit.com/help/contentpolicy, last accessed on August 17, 2016

[70] Reporting links, available at: https://www.reddit.com/wiki/reporting, last accessed on August 17, 2016

[71] How do you choose mods when there are large number of candidates, available at: https://www.reddit.com/r/AskModerators/comments/16zwke/mods_of_larger_subreddits_how_do_you_choose_mods/, last accessed on August 17, 2016

[72] Removing harassing sub-reddits; available at: https://www.reddit.com/comments/39bpam/, last accessed on August 17, 2016

[73] Courtnie Swearingen & Brian Lynch, *We're Reddit Mods, and this is how we handle hate speech*, Wired, August 12, 2015, available at: http://www.wired.com/2015/08/reddit-mods-handle-hate-speech/, last accessed on August 17, 2016

hateful and harassing speech. The second approach, is a rather new method where instead of letting the enraged members to comment in unrelated forums, they choose to cut off any support to these sub-reddits. They isolate them, not provide any resources, or permission to promote their agenda. Therefore, they exist in a small corner, where their effects are not visible to the rest of the Reddit community. Reddit is enhancing its moderation management with plans of introducing a sub-reddit called ModSupport where moderators can talk about the problems they face and the techniques they use.

Moderation of content on social networking and other websites may not always be algorithmic. There are people, who have been outsourced this job of sifting through sites to remove material that may be against the terms of service and standards. Many US companies outsource the moderating part of the job to the Philippines and other developing countries where college graduates spend their days looking at pictures, videos, and text and sort it into differing piles and decide if the content violates the rules and standards of the forum. One of the workers provided an insight into the test he used to determine if the content would violate any terms. He said that one has to ask, “What is the intention (of this post)? The workers have to determine the difference between thought and solicitation.” The CEO of Twitter, mentioned very subtly that the task of moderating content by humans is not logistically possible with the technical and global vastness of the service.

A study conducted in 2014 by Take back the Tech, an organization working on making ICTs safe for women, reported on the effectiveness of the take down mechanism employed by popular social media platforms of Facebook, Twitter & YouTube[74], revealing that none of these widely used pltforms have adequate transparency around reporting and redress mechanisms. Both Twitter and YouTube were given a 'D' grade, meaning there were reluctant attempts or minor actions with no significant results, whereas Facebook did slightly better and managed a 'C' grade, implying that there were previous commitments made to take action or there was effective action in some areas, but lack of complete follow through and serious areas of inaction remain. On the point of simplified and easily accessible reporting mechanism, none of the platforms managed to show that sufficient effort had been made on their part and thereby, received a 'C' grade.

As the role of the platform companies as the place where many modern day discussions expands, the expectations that the users and state actors have from them also see an upswing. The issues are not simple and continuous engagement with various actors, transparency and evolution of community standards are highly warranted to address areas of ongoing challenge for companies.

[74] *Take Back The Tech's Report Card on Social Media and Violence against Women*, 2014; available at: https://www.takebackthetech.net/sites/default/files/2014-reportcard-en.pdf, last accessed on May 2, 2016

6.2 Counter-initiatives of note

The world has become sans boundaries with the increasing Internet penetration and advancement. The issues of hate speech, cyber bullying, violence against women in the virtual world, and other such forms of online intimidation have become a cause of concern all across the globe. This digital overtake has led to the creation of movements that have made commitments and subsequent efforts to ensure that the Internet can function as a safe space. Many of these global movements use the Internet as a platform to organize campaigns, share experiences, make their resources available to a larger population and collaborate as a community to enhance their operations. This section provides a glimpse of such movements that have dedicated themselves to mapping hate speech, creating educational tools to address bullying and online harassment.

6.2.1 Code of Conduct on illegal online hate speech

Following the EU Colloquium on Fundamental Rights in October 2015 on *'Tolerance and respect: preventing and combating Antisemitic and anti-Muslim hatred in Europe'*, the European Commission initiated a dialogue with major IT companies including Facebook, Twitter, Google and Microsoft, in cooperation with EU Member States and civil society, to explore effective means to tackle illegal online hate speech.[75] The terror attacks in Brussels and the increasing use of social media by terrorist groups to radicalize young people lent more urgency to tackling this issue. The Joint Statement of the Justice and Home Affairs Council following the Brussels terrorist attacks underlined the need to step up work in this field and also to agree on a Code of Conduct on hate speech online.[76]

By signing this Code of Conduct, the IT companies committed to make efforts to tackle illegal hate speech online, including by ensuring the continued development of internal procedures and staff training to guarantee that the majority of valid notifications for removal of illegal hate speech is expediently reviewed and access to such content disabled, if necessary. The IT companies will also endeavor to strengthen their ongoing partnerships with civil society organizations who will help flag content that promotes incitement to violence and hateful conduct. The IT companies and the

[75] European Commission – Press Release, *European Commission and IT Companies announce Code of Conduct on illegal online hate speech*, 31st May 2016, available at: http://europa.eu/rapid/press-release_IP-16-1937_en.htm, last accessed on 26th June, 2016

[76] Joint statement of EU Ministers for Justice and Home Affairs and representatives of EU institutions on the terrorist attacks in Brussels on 22 March 2016, 24th March 2016, available at: http://www.consilium.europa.eu/en/press/press-releases/2016/03/24-statement-on-terrorist-attacks-in-brussels-on-22-march/?utm_source=dsms-auto&utm_medium=email&utm_campaign=Joint%20statement%20of%20EU%20Ministers%20for%20Justice%20and%20Home%20Affairs%20and%20representatives%20of%20EU%20institutions%20on%20the%20terrorist%20attacks%20in%20Brussels%20on%2022%20March%202016, last accessed on 26th June, 2016

European Commission also aim to continue their work in identifying and promoting independent counter-narratives, new ideas and initiatives, and supporting educational programs that encourage critical thinking.

Some of the notable public commitments contained in the Code of Conduct are:

- Upon receipt of a valid removal notification, the IT Companies will review such requests against their rules and community guidelines and where necessary national laws transposing the Framework Decision 2008/913/JHA, with dedicated teams reviewing requests.
- The IT Companies will review the majority of valid notifications for removal of illegal hate speech in less than 24 hours and remove or disable access to such content, if necessary.
- The IT companies will provide information on the procedures for submitting notices, with a view to improving the speed and effectiveness of communication between the Member State authorities and the IT Companies, in particular on notifications and on disabling access to or removal of illegal hate speech online.
- The IT Companies will encourage the provision of notices and flagging of content that promotes incitement to violence and hateful conduct at scale by experts, particularly via partnerships with civil society organizations, by providing clear information on individual company Rules and Community Guidelines and rules on the reporting and notification processes.
- The IT Companies rely on support from Member States and the European Commission to ensure access to a representative network of civil society partners and "trusted reporters" in all Member States helping to help provide high quality notices. IT Companies to make information about "trusted reporters" available on their websites.

The IT Companies and the European Commission agreed to assess the public commitments in this code of conduct on a regular basis, including their impact. They also agreed to further discuss how to promote transparency and encourage counter and alternative narratives. To this end, regular meetings will take place and a preliminary assessment will be reported to the High Level Group on Combating Racism, Xenophobia and all forms of intolerance by the end of 2016.

Despite being the first European effort to unify policy on hate speech across the EU, the Code of Conduct has been criticized for delegating tasks to private entities that should be carried out by law enforcement.[77] A joint statement by European Digital Rights (European Digital Rights), a Brussels-

[77] EDRi and Access Now Withdraw from the EU Commission IT Forum discussions, 31st May, 2016, available at:

based civil society organization, and Access Now said in a joint statement that they would withdraw from future discussions, saying that civil society organizations were "systematically excluded" from negotiations over the code of conduct.[78] The joint statement further stated that the 'code of conduct' downgraded the law to a second-class status, behind the 'leading role' of private companies that were being asked to arbitrarily implement their terms of service.

6.2.2 No hate speech movement

This campaign is a part of the Council of Europe's program on Young People Combating Hate Speech Online, which was in operation between 2012-2014. The 'No Hate Speech Movement' is a project against hate speech, racism, and discrimination in the online form of expression. This campaign branches out into national campaigns in all members of the European Union.[79] Through youth participation and co-management, it aims to promote human rights education and media literacy about online hate speech.

In their three preliminary studies, they identified the key issues by analyzing the relation between regulation of hate speech and freedom of expression and devised guidelines to be followed during the movement. Their offline activities in the respective national campaigns, include creative workshops for educational and awareness building purposes and meeting with the stakeholders to focus on context specific issues, especially to target hate speech in their local languages. Their online tools include training online bloggers and activists through education and training modules on how to combat hate speech. One of their major online activity includes creation of a forum called the 'Hate Speech Watch', which is a community sourced database on various incidents of online hate speech.[80]

With focus on awareness and advocacy, the 'No Hate Speech Movement' has compiled their experience and learning into various toolkits for organizing campaigns and training online human rights activist. They have also published a manual called Bookmarks for combating hate speech online through human rights education.[81] They are involved in developing literature for producing counter narratives for hate speech, and though they work in the European context, they aim to create

https://edri.org/edri-access-now-withdraw-eu-commission-forum-discussions/, last accessed on 26th June, 2016

[78] Amar Toor, *Facebook, Twitter, Google and Microsoft agree to EU hate speech rules*, The Verge, 31st May 2016, available at: http://www.theverge.com/2016/5/31/11817540/facebook-twitter-google-microsoft-hate-speech-europe, last accessed on June 26, 2016

[79] Welcome to the No Hate Speech Movement- Campaign of young people for human rights online; available at: http://www.nohatespeechmovement.org/campaign, last accessed on June 26, 2016

[80] No Hate Speech Movement, Methodology, available at: http://nohate.ext.coe.int/The-Campaign/Methodology2, last accessed on July 8, 2016

[81] No Hate Speech Movement, Bookmarks, available at: http://www.nohatespeechmovement.org/bookmarks, last accessed on July 8, 2016

modules and material that may be used internationally.[82]

6.2.3 Rabat Plan of Action

On 21 February 2013, the UN Office of the High Commissioner for Human Rights (OHCHR) launched the Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence. This was an effort to draft guidelines that would be helpful in maintaining a balance between Article 19 of the ICCPR, which provides for freedom of expression, and Article 20, which prohibits incitement of discrimination, hostility or violence.[83]

The goal behind this initiative was to conduct a comprehensive assessment of the national and regional legislation, jurisprudence, and implementation of policies pertaining to hostility, violence, or incitement to discrimination, with an emphasis on religious activities. They intended to analyze these policies while balancing the respect for freedom of expression as protected in the international human rights law.[84]

The Rabat Plan of Action was the outcome of four regional expert workshops hosted by the UN OHCHR in Austria, Kenya, Thailand, and Chile during 2011. At each of these consultations, experts and stakeholders discussed the constituents of 'incitement to discrimination, hostility or violence based on national, racial or religious grounds as described in international human rights law' and the methods of balancing it with Article 19 and 20 of ICCPR. In October 2012, at a final meeting in Rabat, Morocco, the OHCHR agreed on this plan of action.[85]

It was agreed in Rabat that there is an absence of domestic legislations that prohibit incitement to hatred. If there is a legislation, the terminology is not always in consonance with Article 20 of ICCPR, or at times has heightened restrictions on freedom of expression. It is recommended that firstly, there should be domestic anti-discrimination legislation that includes preventive and punitive action to effectively combat incitement to hatred. These legislations should have robust definitions of 'hatred, incitement, hostility, violence', etc., for clarity. The three-part test for restriction on freedom of expression i.e. legality, proportionality, and necessity should apply even to the offenses

[82] No Hate Speech Movement, News, available at: http://nohate.ext.coe.int/News, last accessed on July 8, 2016

[83] International Justice Resource Center, *UN launches the Rabat Plan of Action*, February 25, 2013, available at: http://www.ijrcenter.org/2013/02/25/un-launches-the-rabat-plan-of-action/, last accessed on September 23, 2016

[84] Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility, or violence- Conclusions and recommendations emanating from the four regional expert workshops organized by OHCHR, in 2011, and adopted by experts in Rabat, Morocco on 5 October, 2012; available at: http://www.ohchr.org/Documents/Issues/Opinion/SeminarRabat/Rabat_draft_outcome.pdf, last accessed on September 23, 2016

[85] Ibid.

for incitement to hatred. These restrictions should be narrowly defined and targeted to achieve a specific result. It also elaborates on a threshold test for speech that are offenses under criminal law. This six part test considers: the context of incitement to hatred, the speaker, intent, content, extent of the speech, and likelihood of causing harm.[86]

6.2.4 Take back the tech! (TBTT)

An initiative of the Association for Progressive Communications (APC), started in 2006 as a series under APC's Women's Rights Program to create awareness on how ICTs are connected to violence against women. The idea of this campaign is to create an understanding and awareness that the gender disparities that exist offline are also prevalent online and women are increasingly becoming targets of cyber stalking and digital voyeurism.[87]

The core objective of this campaign is a call to everyone, especially women to reclaim technology for the fight against violence against women.[88] Campaigners work for the right to define, access, use and shape ICT for its potential to transform power relations toward equality.[89] TBTT believes that the future of power relations on ICTs, whether they amplify or destabilize will depend on how largely or closely the development of this discourse is monitored by the public.[90]

TBTT celebrates the contribution of women in the ICT sector, shares information, creates collective knowledge, engages in capacity building and education, and creates a network in the community by committing to solidarity.[91] As a part of a daily campaign, the users share online resources for better security in various regions of a country, map incidents, start social media discussions and connect their online and offline activism by sharing lectures, leading workshops and strengthening participation with online mobilization.[92]

Local initiatives of TBTT have enjoyed significant presence in countries such as Brazil, India, Pakistan, DRC, Uganda, the Philippines, Mexico and Uruguay.[93] These local campaigns are known for advancing action and advocacy in their specific regional contexts. Till date, TBTT has translated their information materials to Malay, Czech, Spanish and Portuguese.[94]

[86] Ibid.

[87] Association for Progressive Communications, *Take Back The Tech!*, available at: https://www.apc.org/en/node/2949/%29, last accessed on November 3, 2016

[88] Ibid.

[89] Ars Electronica, *Take Back The Tech!*, available at: http://prix2014.aec.at/prixwinner/13165/, last accessed on November 3, 2016

[90] Take Back the Tech, Know More; available at: https://www.takebackthetech.net/know-more, last accessed on November 3, 2016

[91] Ibid.

[92] Supra. 85

[93] Ibid.

[94] Supra. 83

In the initial period of 2006-2008, TBTT had no specific funding, but focused on creating content and building a platform for collaboration. But, with the significant development of their campaigns, they received funding from 2009-2011 to support work to end online violence against women in twelve countries in Latin America, Africa and Asia, and from 2012 to 2015, in seven countries in the same regions.[95] Apart from being a forum for discussion on Violence against women, TBTT also responds to alerts regarding International Women's Day and the "Stop Cyberbullying" initiative by blogger-networks.[96]

6.2.5 Umati

Umati is a media monitoring project dedicated to understanding and recording incidents of hate and dangerous speech in the online space of Kenya. The 2007-08 Kenyan elections were marred with incidents of riotous violence resulting in the death of as many as 1200 people. Hate speech was identified as one of the major reasons behind these riots.[97] Umati was born in 2012 with the following objectives, out of concern that hate speech may incite further violence in the 2013 elections:[98]

In the Phase I of Umati (2012-2013), the team monitored social media sites, online blogs, news reports, comments, status submissions, tweets for incidents of hate speech. These were then analyzed in accordance to the framework created by Prof. Susan Benesch for detecting dangerous speech. Subsequently, Umati categorized the collected incidents into the categories of offensive speech, moderately dangerous speech, and extremely dangerous speech. This phase was carried out manually with all human process to ensure accuracy. In Phase II, starting 2013, Umati has rolled out an automated mechanism based on their database of more than 7,000 incidents from Phase I. They built a software, Umati Logger, which will collect the required data and classify it using the Machine Language (ML) and Natural Language Processing (NLP) algorithms. The software analyses the comments/text, follows a procedure to create a frequency table of words.[99]

Subsequent to Phase I, the Umati project developed a workable definition of online dangerous speech that is built from the Benesch framework and the findings from Phase I. This definition has

[95] Supra. 85

[96] Supra. 83

[97] Emmanuel Amberber , *Umati: Kenyan platform to fight online hate speech with NLP and machine learning in Africa*, Your Story, September 25, 2014, available at: http://yourstory.com/2014/09/umati-hate-speech/, last accessed on September 22, 2016

[98] Nanjira Sambuli, *Online dangerous speech monitoring in Kenya: Umati Project's findings from January- November 2013*, iHub, June 27, 2014, available at: http://www.ihub.co.ke/blogs/19407, last accessed on September 23, 2016

[99] Leo Mutuku, *Building Swahili stop words corpus for computing*, iHub, April 7, 2014, available at: http://www.ihub.co.ke/blogs/18374/building-swahili-stop-words-corpus-for-computing, last accessed on September 23, 2016

three components; it is targeted at a group of people and not a single person, may contain one of the hallmarks/pillars of dangerous speech, and contains a call for action.[100] Phase II created a dictionary for 'Swahili' words that could be potentially dangerous speech. They now aim to make a similar web dictionary for other African languages.

VII. Indian stakeholders

In India, a number of organizations have invested their resources in studying, generating awareness, working with the policy makers as well as the grassroots to find solutions for the increasing cases of online bullying and harassment in our country. These are mostly civil society actors that are involved in making the digital space a safe platform for everyone, especially women, children, and socio-economically, or sexually marginalized communities. Apart from civil society, there are certain corporate entities, such as Intel Security that have conducted studies on cyber bullying, specific to the Indian context. This chapter will provide a landscape of organizations and groups that are committed to understanding the changing concerns regarding safety on the Internet.

7.1 Centre for Cyber Victim Counseling

This is a non-governmental and not for profit organization that is committed to helping victims of cyber crime in India. Their focus is on helping the victim understand the nature of crime, legal action that can be taken for it, and coping up with the trauma of the event. They provide assistance and help to people who have suffered or encountered cyber harassment, hate speech, cyber stalking/bullying, identity theft, victims of social networking harassment, women and teen victims. They work through an online form submission portal wherein a complainant can submit their query/complaint and will be contacted by the Counseling center. Over the years, this organization has done several surveys, reports, and drafted policy guidelines including a recent 2015 research report on the harassment over WhatsApp in rural and urban India. They also published a report in 2013 about the Misuse of Internet by Semi-urban and Rural Youth in India and in 2010 about the Cyber Victimization in India.

7.2 Digital Empowerment Foundation

DEF is an organization that works at the grassroots, understanding the improvement that Internet and digitalization can bring in the socially and economically marginalized communities. Their mission is to increase Internet usage and provide access to wider knowledge, digital literacy and tools to create not only awareness, but also focus on development of education and micro institutions. They aim to bridge the digital divide and the technology gap at the ground level and

[100] Supra. 98

integrate it with their demands and solutions. Along with Google India, DEF began an initiative called the 'Digital Literacy, Safety, and Security' in 2013. Through workshops, it aims at promoting and imparting means and tips to ensure safety and security of users when browsing the Internet.

Digital Empowerment Foundation has been a member of the Association for Progressive Communications since 2009 and jointly submitted the Universal Periodic Review Report in 2012. This report has a specific mention of women´s rights and recommends to Indian Government to adopt a rights based approach in reviewing the Information Technology Act: *“Women’s use of the Internet shows that Internet content is regulated by four factors: access and infrastructure, law and policy, markets and economic forces and culture and social norms. Research on female use of the Internet in India reveals that these four factors also affect women’s access to and use of the Internet and that the Internet has significant implications for women’s communication rights and sexuality rights”.*

7.3 Intel Security India

Intel Security works to promote and educate the public about security and safety while using the Internet. They work with policy makers, critical infrastructure sectors, global governments to further their goal of fighting cybercrime and cyberbullying. It is one of the founding members of the National Cyber Security Alliance – a US-based non-profit organization seeking to promote cyber security awareness for home users, small and medium size businesses, and primary and secondary education. Intel Security initiated a Digital Safety Program[101] that is designed to provide guidance about safely accessing the Internet to students, families, and seniors. The means employed under this program include digital safety tips, insightful blogs, research on technology usage of tweens, advice on how to protect smart-phones and mobile communications. Apart from this program, Intel Security has other digital initiatives that pursue the goal of research and awareness in these avenues. They have developed a cyber educational module called 'bCyberwise' with the help of Life Education, which is targeted at middle-primary school students and includes topics on being responsible online behavior, how to keep personal information personal, strategies to deal with cyber bullying and many more.[102] They have conducted surveys and research studies in 2014 and 2015 on cyberbullying and social media usage patterns of teens, and tweens specific to the Indian context, the latter of which found that 43% of the children active on social media claim to have witnessed cruel behaviour on social networks, while 52% of the children indicate that they have

[101] Intel Security Digital Safety Program, available at: http://www.mcafee.com/in/about/intel-security-digital-safety.aspx, last accessed on October 5, 2016

[102] bCyberwise Monster Family App, available at: https://www.lifeeducation.org.au/teachers/item/51-bcyberwise-monster-family-article-in-for-teachers, last accessed on October 5, 2016

bullied people over social media themselves. Of these, the study found that 27% made fun of others, 24% called someone fat or ugly or made fun of other physical appearances, and 23% tagged mean pictures. Reasons cited for cyberbullying another child were because the others were mean to them (49%) or they just don’t like the other person (28%).[103]

7.4 Internet Democracy Project

Internet Democracy Project (IDP) analyses the role of Internet in changing the social structures in a democracy. Through research, advocacy and debate, IDP seeks to unearth both the changes wrought by technology to democracy as we know it

IDP focuses its resources on studying issues surrounding gender, free speech, and censorship in a digital age, the use of hate speech in online forums, the implication of various laws and policies already in place to regulate expression on the Internet. It also produces periodic research reports and articles and holds discussions as a means to generate awareness and gauge the perspectives of relevant stakeholders. Two of IDP's publications are of particular interest to the topic of online harassment:

- *'Don't Let it Stand!' An Exploratory Study of Women and Verbal Online Abuse in India:*[104] The purpose of this study is to explore gender-based abusive speech online. By speaking with Indian women who are active users of social media, the study considers instances of abuse, how and why they occur, what forms they take, and if and how women strategize in order to address this verbal abuse or hate speech. An exploration of these issues also considers the extent to which women Internet users consider legal recourse to be a useful strategy.

- *Keeping Women Safe? Gender, Online Harassment and Indian Law:*[105] Published as a complementing paper to the above-mentioned study, this report begins with an examination of Indian obscenity laws, which can potentially be mobilized to fight gendered online abuse. The report then analyzes in greater detail other provisions in the law that can be drawn on to specifically address the verbal abuse of women online, and concludes by putting forward a number of possible legal amendments that have emerged over the course of our research as

[103] Intel Security India, *Press Release on Teens, Tweens and Technology Study*, October 27, 2015, available at: http://apac.intelsecurity.com/digitalsafety/wp-content/uploads/sites/7/2015/10/Intel-Security_India_Press-Release_TeensTweensTech_271015.pdf, last accessed on October 5, 2016

[104] Anja Kovacs, Richa Kaul Padte, Shobha S V, *An Exploratory Study of Women and Verbal Online Abuse in India*, April 2013, available at: https://internetdemocracy.in/wp-content/uploads/2013/12/Internet-Democracy-Project-Women-and-Online-Abuse.pdf, last accessed on November 1, 2016

[105] Richa Kaul Padte, *Keeping Women Safe? Gender, Online Harassment and Indian Law*, Match 2013, available at: https://internetdemocracy.in/wp-content/uploads/2013/04/Internet-Democracy-Project-Gender-Online-Harassment-and-Indian-Law.pdf, last accessed on November 1, 2016

potential ways forward to provide better protection to women who face abuse online.

7.5 IT for Change

IT for Change is an NGO located in Bengaluru that aims to promote the effective integration and use of ICTs as a means of socio-economic change in the developing world. Through research, field work and advocacy, they work - in the thematic areas of development and ICTs, gender, education, technology governance among others. As a stakeholder, IT for Change has done various projects and reports, in the last few years with the aim of sensitizing the digital environment. They contributed to a two-year action research project conducted by International Development Research Centre (IDRC), Canada to study the role of digital and ICTs in strengthening the participation of women in governance processes in the global south. The places chosen for this study were Rio de Janeiro (Brazil), Mysore (India) and Cape Town (South Africa). Furthermore, IT for Change has also been involved with the United Nations Economic and Social Commission for Asia and the Pacific (UNESCAP) to build a framework for their research on e-governance and gender equality in five countries of the Asia-Pacific region. Recently, they also conducted a three-day short course on 'Rewiring Women's Rights Debates in the Digital Age' in New Delhi. This course brought together a variety of participants, ranging from NGO leaders as well as young researchers working in the areas of social justice. This course was aimed at initiating conversations about difference in the gender rights discourse with the onset of the digital age.

7.6 JaagoTeens

This NGO, as the name suggests is targeted at teenagers and children. It was started by three Indian mothers for raising awareness about Internet safety and cyberbullying among adolescents and young adults. Their methodology includes conducting campaigns and workshops in schools and educating through writing various articles and blog posts. They help children understand the concepts of known and unknown on the Internet, how to search for authentic information, proper email and blog usage, proper etiquette on the net, along with giving them tips on how to stay away from inappropriate content. This grassroots project, established in New Delhi is working on a small scale, targeting the young generation, and giving them adequate understanding on how to be secure online.

7.7 Point of View

Point of View is a Mumbai-based non-profit organization that aims to amplify the voices of women in India and remove barriers to free speech and expression. Its work spans multiple forms of media, art and culture both offline and online, and has five key program areas viz. placing the broad concept of gender in the public domain; putting forward the realities of women in sex work as they see them; highlighting marginalized issues of gender, sexuality and rights; enabling women to speak out about and prevent domestic violence; and building the capacities of grassroots women to express themselves through media platforms.

Notably, POV is the Indian partner of EroTICS (Exploratory research on Information and Communication Technologies) – a project spearheaded by the Association for Progressive Communications in 2009 as an exploratory study to understand the use of Internet for advancement of sexual rights, and gender rights. Their discussions engage sexual health workers and activists, policy makers, advocates and other stakeholders to understand the varied perceptions of 'harm', privacy and security on the digital space. Now in its second phase of operation, EroTICS aims to build a network of advocacy around Internet and sexual rights, which would be able to collaborate stakeholder expertise and respond adequately to the instances for participation of sexual minorities on the Internet.

As part of the EroTICS project, POV periodically publishes articles and reports on topics such as women, sexuality and the Internet, gendered abuse online, and comparing online abuse of women with street harassment. POV is also responsible for the establishment of the Internet Democracy Project.

7.8 SFLC.in

SFLC.in is a Delhi-based not-for-profit legal services organization that works to protect citizens civil liberties in the digital world. Freedom of expression online has been one of its core areas of work since its inception back in 2008, and it has worked extensively on the campaign against the previously mentioned Section 66A of the Information Technology Act, which allowed for the incarceration of innocent citizens over online content that is subjectively and arbitrarily deemed offensive. The organization has also undertaken significant work around the topic of intermediary liability, and spearheaded the filing of a writ petition before the Supreme Court of India in the year 2013 that led to the problematic Information Technology (Intermediaries Guidelines) Rules, 2011 being read down in the interest of much needed clarity. Prior to being read down, the Rules forced intermediaries such as Google and Facebook to play the role of adjudicators and take down content on the receipt of notices to the effect from any member of the general public. SFLC.in was part of the batch of litigants that moved the Supreme Court of India to read down the Rules so that content take downs could occur only on the receipt of a court order or Government directive, thereby limiting the scope for private censorship of legitimate free speech and relieving intermediaries of adjudicatory functions that they were ill-equipped to handle. Aside from these, SFLC.in frequently organizes crypto-parties, where those interested are taught how to better incorporate anonymization tools and encryption technologies into their online behavior so as to protect themselves from undue surveillance and from other online attacks.

VIII. Observations and recommendations

Over the preceding pages, we have examined popular definitions and key terms associated with online harassment and briefly looked at its potential to cause real-world violence. We heard from individuals who have either faced online hate themselves or have worked in close proximity with the issue. Prominent state and non-state responses to the issue were examined, and we have seen a few global campaigns and domestic stakeholders that do great work towards identifying and limiting this problem.

It is clear that targeted and sustained expressions of hate and intolerance online are capable of dealing great damage to individuals in isolation and society as a whole. As the boundaries between online and offline worlds continue to blur at alarming rates, both the benefits and perils that surface in the digital age can cause lasting effects both in the long and short runs. While India's legal machinery does contain provisions that would enable perpetrators of online hate, intolerance and harassment to be held accountable, it is seen that the actual enforcement is burdensome for the complainant and spotty. What cannot be ignored is the abuse of such laws and the desire of several victims to avoid resorting to law enforcement other than under the most extreme of circumstances. Law enforcement officials – even those representing dedicated cyber crime cells – appear under-prepared to register and handle complaints relating to online hate. Granted, the very nature of the Internet, where complete anonymity is almost always an option, might present several complex and tangible challenges to effective resolution of grievances. It is imperative that the law enforcement functionaries of today be equipped to provide sufficient guidance to victims in the very least, as opposed to expressing their helplessness.

Given the global nature of Internet, combating online extremism presents enormous difficulties, and it cannot be done only within the borders of individual countries. Therefore, international cooperation is essential, and the work of different international associations and networks should be encouraged. Attention should also be paid to educating people about existing mechanisms for combating online hate speech, so that each Internet user would be aware of the power they have to make a difference.

8.1 User-level safeguards against online harassment

In the meanwhile, below are a few measures that would help minimize the possibility of becoming a target for online harassment. While most of these steps are fairly intuitive and common-sensical, it would nevertheless do well for those at risk of being targeted to internalize this list of do-s and don't-s so that their opinions and convictions expressed online don't open the floodgates to rampant abuse.

- *Thoroughly screen the personal information shared online* – Be very careful about what personal information you make publicly available, and refrain from providing any information apart from that which is absolutely essential for its purpose. Do not feel obligated to fill out all fields when registering online, and wherever possible, avoid providing identifying information such as birth-date and place in required fields. It is very easy to glean information about where you live, the places you love to go to in your area and the people you care about from posts and pictures.
- *Consider dedicating an email ID for social-media purposes* – Create and maintain an email ID which is dedicated solely for supply while signing up and/or using social media services. Do not use this ID for personal communication purposes. This will help avoid spam and your personal email will not be revealed if the online service doesn’t have a good privacy practice.
- *Avoid uploading photographs that identify you or your location* – Do not upload photographs that identify either your person or your location to their viewers.
- *Use a pseudonym* – Rather than using your real given name, use a gender and age neutral pseudonym, if anonymity is relevant in your online activities.
- *Keep a tab on information others post about you* – Keep in mind that personally identifiable information need not always originate from you. Periodically review information about you that others have posted so as to ensure that none of it may lead to your being identified by unwanted parties. Let your friends and family know your concerns about privacy and help them learn better privacy settings.
- *Run Internet searches on yourself* – Regularly monitor where you appear online. If you find unauthorized information about yourself online, contact the relevant website moderator to request its removal.
- *Use strong passwords and change them periodically* – Use a mix of different types of

characters such as alphabets, numbers and special characters to make your password harder to crack. Stay away from obvious dictionary words and combinations of dictionary words, and avoid using common dates such as your birthday as the digits in your password. Also be sure to periodically change the password to minimize the risk of its being compromised.

- *Review your service providers' privacy policies* – Services such as Facebook change their privacy policy all the time, so it is a good idea to check your privacy settings to make sure you are sharing the information you want to share with people you trust and not the general Internet public. Some sites have options for you to test how your profile is being viewed by others – test and make sure you only reveal what is absolutely necessary.

If on the other hand, you find yourself at the receiving end of targeted online harassment campaigns of any scale or complexity, the following steps would help contain the situation and ensure that you have the necessary material at your disposal to see any action initiated in response to completion.

- *Record all communications with the perpetrators* – If you are receiving unwanted contact, make clear to that person that you would like him or her not to contact you again. Depending on the harasser, engagement with the person can escalate or cease, so if you consider contact appropriate and necessary, do so once and document it. Do not edit or alter the communications in any way. Try using print screens, especially if the harassment is happening in real-time.
- *Report incidents to the concerned service providers* – Many online service providers, especially social media platforms, have detailed protocols and procedures in place for reporting and resolving complaints related to objectionable content, including hateful, intolerant and harassing contact. Familiarize yourself with these protocols and do not hesitate to use them wherever necessary.
- *Block the perpetrators* – Blocking your abusers, as almost every major online speech platform allows, is a very effective way to put and end to the abuse once and for all. While this might not be particularly feasible when the abuse comes in high volumes, it is a foolproof measure when the perpetrators are limited in number.
- *Approach law enforcement* – This would ideally be a last resort for when there are real threats to your physical safety, but do not hesitate to approach the relevant LEA to register your complaint. There always are legal remedies available to such issues, though actually availing these might be more complicated than it should be.

- *Seek help from social media influencers* – By leveraging their extensive networks, social media influencers will be able to invite public attention to your issue. Seeking help from them can rally large numbers of people in your support and initiate effective counter-narratives.
- *Record all communications with service providers and law enforcement* – Even though the instinct might be to delete harassing communications, these records carry significant evidentiary value. Back up these communications on another computer or removable memory stick or external hard drive.
- *Seek support from friends and family* - Being harassed – online or offline – is a traumatic experience and support from family and friends will help you cope. Also check what they are revealing about you and their relationship with you in their online spaces, albeit inadvertently. Highlight the problem in public forums where possible, as not only will this generate greater support, but it will also go towards raising awareness about the issue in general and bringing it under public scrutiny.

8.2 Draft best practices to limit online harassment

Considering how the most ideal responses to harmful speech would stem at the non-state level as previously mentioned, we further propose that self-regulatory measures be considered for adoption by online intermediaries that function as speech platforms in any capacity, where users have the option of creating and publishing content without pre-filtrations. We re-emphasize that any effort at limiting even harmful speech will inevitably come with challenges as to avoiding collateral limitations on legitimate speech and consequently, this list of best practices should be seen strictly as proposals that will require substantial deliberations before being formalized. We hope the following measures, inspired in part by the European Union's previously mentioned code of conduct against illegal online hate speech, serve as a starting point for an effective dialogue in this regard:

- Have in place rules that prohibit hateful, disparaging, and harassing content on intermediary networks; rules must be clearly articulated and designed for easy consumption; include illustrative examples for each category of prohibited content
- Generate awareness within user community on prohibited content; notification systems, promotional banners etc. could be leveraged for the purpose
- Enable easy and accurate reportage by users and third-parties; include easily identifiable “report” buttons; provide adequate opportunities to substantiate why content must be

removed;

- Have clearly defined review processes prescribing (where possible) objective standards for determining permissibility; refer to applicable national laws
- Deploy dedicated teams to review and disable content; provide periodic training to review teams on efficient identification and disablement;
- Review reports and disable content within a prescribed time frame (24/48/76 hours)
- Provide opportunities to creators of disabled content to justify themselves; include provisions for timely restoration of disabled content and reinstation of terminated accounts
- Share best practices within stakeholder community; contribute to building effective multi-stakeholder norms for tackling prohibited content
- Liaise with law enforcement; aid in investigation of reported offenses in consonance with established legal procedures
- Work with other stakeholder communities; engage with civil society organizations and academia on awareness generation; conduct trainings/workshops for law enforcement officials on reportage mechanisms so as to facilitate effective handling of complaints
- Promote counter-speech; invite counter narratives from public figures; offer incentives; conceptualize additional means to promote counter-speech.

8.3 Conclusion

There can be no doubt that "words with the effect of a blow," aimed to wound or to harm, are appropriate subjects of regulation in any legal system. Not only is it reasonable to hold speakers responsible for the intentional harms caused by their speech, it is reasonable to expect those who amplify and transmit that verbal form of violence into the spaces that have always been privileged, that have always provided security against hateful communication, to exercise responsibility as well. We are now building through the Internet, a city as wide as humankind itself. How we carry our civility into that new space, how we learn to deal with fighting words in the net, determines whether we are also building a better human civilization.

To say that online harmful speech in general, and online harassment in particular, are complex areas of research would be an understatement. Blurry boundaries and overlaps are characteristic of this domain, and narratives are made even more complex by the lack of consensus at multiple levels, including terminologies, definitions, and desired responses. This report is by no means comprehensive in its coverage of the core issue, and we strongly caution readers against treating it as such. This is but the first in a series of studies that SFLC.in intends to undertake in this domain, and we expect our own understanding to improve with each iteration of output. In the meanwhile, we have constituted a Working Group on Online Harassment, comprising industry and civil society actors and academicians collectively working to better understand and tackle this increasingly important issue. We invite our readers to join us as we attempt to build a sustainable dialogue around online harassment, as participatory and result-oriented initiatives are the need of the hour.

Annexure 1

Twitter's feature additions to combat online harassment

Until November 2016, there were three primary ways in which Twitter users could choose to respond to online harassment encountered on the platform:

- *Mute accounts* – Muting accounts will make all tweets from the muted accounts disappear from the users' Twitter timelines. These accounts are still able to follow each other and exchange direct messages, and the users will still be notified of replies/mentions from muted accounts.[1]
- *Block accounts* – Once a user blocks another user's account, the blocked account will no longer be able to follow them, view and search for their tweets, send them direct messages, and tag them in photos among other things. If a blocked account visits the profile of the user who blocked them, they will see they have been blocked. Tweets from the blocked account will also no longer appear in the user's timeline, though tweets from others accounts that the user follows, where the blocked account is mentioned, and tweets mentioning both the user's account and the blocked account will still be visible to the user.[2]
- *Report tweets and/or accounts* – Users can report particular tweets and entire accounts directly to Twitter, if they have reason to believe that these accounts are in violation of Twitter's content policies, including the Twitter Rules. Once a such a report is initiated, Twitter may ask the user to provide more information as to why the reported tweet and/or account is in violation of Twitter policies. Submitted reports are reviewed by dedicated review teams at Twitter, and appropriate action is taken, which may include removal of reported content and permanent suspension of reported accounts depending on the severity of the violation.[3]

On November 15, 2016, Twitter announced a set of new features designed specifically to address the growing trend of online abuse, to be made available alongside the existing mute, block, and report options.[4] Most importantly, it was announced that Twitter users would be given the option of muting not just entire accounts, but also notifications of tweets mentioning them and containing particular words, phrases, usernames, emojis, and hashtags. Tweets containing the muted attributes

[1] Muting accounts on Twitter, available at: https://support.twitter.com/articles/20171399, last accessed on November 16, 2016

[2] Blocking accounts on Twitter, available at: https://support.twitter.com/articles/117063, last accessed on November 16, 2016

[3] How to report violations, available at: https://support.twitter.com/articles/15789, last accessed on November 16, 2016

[4] Progress on addressing online abuse, November 15, 2016, available at: https://blog.twitter.com/2016/progress-on-addressing-online-abuse, last accessed on November 16, 2016

will be removed from the user's notifications tab, push notifications, and SMS and email notifications. However, such muted tweets will continue to appear in the user's timeline and via search. It was also announced on November 16 that users would be given “a more direct way” to report content that violated Twitter Rules by targeting people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease, though it was not immediately clear how this would be different from the existing report mechanism. Furthermore, all Twitter support team staff have been retrained on the Twitter policies, including through special sessions on cultural and historical contextualization of hateful conduct, and an ongoing refresher program was introduced. Twitter also announced that its internal tools and systems have been updated so as to deal more effectively with hateful and abuse conduct, though no details were available at the time of writing this report as to the nature of these updates.

One final protective feature worth noting, which existed well before the November 2016 feature additions, is the option given to Twitter users to “protect” their tweets i.e. to prevent non-followers from viewing and interacting with their tweets. All tweets made by a user are “public” by default, and any Twitter user is able to view and interact with the public tweets of another user. Once a user chooses to protect their tweets through a toggle available under their privacy settings however, the following will changes will apply:[5]

- Each new follower will need to be manually approved by the user. Followers will also be unable to re-tweet or quote their tweets.
- Their tweets will be visible only to their followers, and the tweets will not appear in third-party search engines like Google and Bing. Protected tweets are only searchable by the user and their followers on Twitter.

Though this seemingly does not prevent potential harassers from identifying and mentioning their targets in tweets, protected tweets do offer an extra layer of security that lets users minimize their public exposure and by extension, the odds of being targeted for things they share on Twitter. All in all, Twitter has demonstrated a gradual awakening to the widespread prevalence of bullying and harassment on the platform, and it is seen taking incremental and public facing measures to limit such instances. However, the reporting tools available to users are far from comprehensive, and leaves much to be desired. Long turn-around times when it comes to acting on content reports has also been a persistent complaint. Whether the latest round of feature additions make any appreciable difference in this regard remains to be seen.

[5] About public and protected tweets, available at: https://support.twitter.com/articles/14016, last accessed on November 16, 2016