

Encouraging Counter-Speech by Mapping the Contours of Hate Speech on Facebook in India

Maya Mirchandani

with Ojasvi Goel and Dhananjay Sahai

Encouraging Counter-Speech by Mapping the Contours of Hate Speech on Facebook in India

Maya Mirchandani

with **Ojasvi Goel** and **Dhananjay Sahai**

© 2018 by Observer Research Foundation

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means without permission in writing from ORF.

Encouraging Counter-Speech by Mapping the Contours of Hate Speech on Facebook in India

ISBN: 978-93-87407-91-6

Printed by:
Mohit Enterprises

INTRODUCTION

In July 2017, Facebook users in India crossed 240 million, pushing past its user base in the US and making India the largest audience for the social platform. Globally, India accounts for over 10 percent of Facebook's users, making the platform a unique place to create, engage in, inform and influence opinions and debate in the country.

The exponential growth in the use and popularity of Facebook is a result of its ability to facilitate positive interaction with friends and family and larger communities, and disseminate information efficiently. Despite its benefits, however, Facebook is also being misused by a small but growing number of people to engage in abusive or hateful speech targeting individuals or communities. Such speech often contains provocations that lead to violence or extremism. To effectively counter violent extremism and initiate an alternative narrative, it is imperative to understand the extent to which hate speech occurs.

Efforts at Countering Violent Extremism (or CVE in internationally accepted terminology) online have become an important focus for all social networks. CVE targets violent, extremist ideologies at their core, tackling them via alternate narratives that focus on peace-building through community interaction. It has thus become an invaluable tool to supplement counterterrorism strategies worldwide.

To identify effective counter-speech on the platform, the Observer Research Foundation (ORF) conducted a study, with support from Facebook, to analyse posts and comments on prominent public pages posting in India. These pages belong to mainstream news organisations, community groups, religious organisations, and prominent public personalities. In a section titled "Controversial, Harmful and Hateful Speech" in its "Community Standards," Facebook details its efforts to make its platform a safe and respectful place for engagement.¹ These standards, according to Facebook, have evolved over the years using feedback from

groups that have faced discrimination on the basis of religion, sexual identity and gender, and have dealt with violence in the form of posts and comments or video and images.

Social media companies are constantly under pressure to identify hate speech and act quickly to remove it from their platforms.² While there is no global consensus on what constitutes hate speech, in an article on its newsroom page on 27 June 2017,³ Richard Allen, Facebook's Vice President for Public Policy, reiterated the platform's definition as anything that directly attacks people based on their "protected characteristics" of race, ethnicity, religious affiliation, sexual orientation, national origin, gender, gender identity, disability or disease.⁴ However, there is uncertainty about unequivocally defining a post or comment as hate speech: often, the words are ambiguous or the context, unfamiliar.

Social media users often cite examples of hateful speech, accusing those who disagree with their views of "trolling," and spewing verbal abuse and threats against them. According to Facebook, it removes approximately 288,000 posts per month globally, after verifying the violations that are reported.

For this study, it was important to first arrive at an understanding of what constitutes hate speech, especially online, where communication thrives on the right to 'freedom of expression' and the protection it can offer, particularly when speech borders on abuse and incitement. The internet has revolutionised and democratised communication, making all speech (including hate speech) travel much faster and across larger distances, and thereby reaching a broader audience. The proliferation and increasing penetration of the internet through mobile connectivity in India has made the debate on hate speech far more pressing. The internet is also challenging systems of social interaction: much of the extreme content is posted online privately, often without the sanction or knowledge of family or community elders who could play a crucial role in countering violent extremism.

DEFINING ‘HATE SPEECH’

Defining ‘hate speech’ is particularly difficult because it is uniquely tied to the impact of the speech itself: there may be situations where the intent and vocabulary is not explicitly hateful but the implications are. Understanding the implications require an understanding of the context in which the speech takes place, which in turn needs a more nuanced approach to the study of hate speech. The attempt in this study is threefold:

1. Determining the intent and implication of the post/comment through tone and language;
2. Addressing differences in determining what constitutes hate speech around the world; and
3. Determining the trans-boundary nature of hate speech.

For the purposes of this study, “hate speech” refers to expressions that advocate incitement to harm—discrimination, hostility, violence—based upon the targets being identified with a certain social or demographic group. It includes speech that advocates, threatens or encourages violent acts.

The difference in the understanding and definitions of what constitutes hate speech around the world has often led to confusion when it comes to a legal treatment of it, as cases come up. Should the government regulate, or should there be greater self-regulation? These are challenging questions for governments, organisations and individuals, especially because speech on online platforms such as Facebook crosses geographical boundaries. For instance, this study shows that users posting hateful commentary on the issue of Kashmir were often based across the border in Pakistan. Thus, any approach to solving these issues must keep in mind complex geopolitical realities.

In a socially networked world, where comment is free and reactions instant, lines between violent personal abuse and/or speech inciting violence against a community or group are becoming increasingly blurred,

especially when a community is targeted through an individual, making the definitions of hate speech broader and open to interpretation. Hate speech can include words that are insulting to those in power and derisive or derogatory of prominent individuals. At critical times, such as during election campaigns, the concept of hate speech may be prone to manipulation. Accusations of fomenting hate speech may be traded among political opponents or used by those in power to curb dissent and criticism. Personal abuse, hate speech and violent extremism often exist in the same ecosystem, with one feeding into the other, not necessarily in a linear way.

Both, ensuring freedom of expression for its citizens and countering violent extremism are significant priorities for any state. Hate speech falls into a space between the two; while a heavy-handed approach compromises free expression, unfettered (free) speech can result in a wider spread of extremist messaging. Given these realities, the extremely fine line between protecting the fundamental freedom of expression and shutting down hate speech often gets blurred.

In the US, where Facebook is headquartered, free speech is protected as an absolute and fundamental right under the First Amendment. Within such a framework, even the most offensive forms of expression—underscoring xenophobia, racism and religious discrimination—are often protected, unless they result in direct violence against an individual or a group.

By contrast, several European nations place value on principles of “civility and respect”⁵ and personal honour, as argued by James Whitman, a professor of comparative law at Yale University. Germany’s Basic Law of 1949, for example, guarantees human dignity in its First Article, and “protection against insult” is a right under German law.

A UNESCO report places hate speech in a “complex nexus with freedom of expression, individual, group and minority rights, as well as concepts of dignity, liberty and equality.”⁶ The United Nations Human Rights Council in its Report of the Special Rapporteur on the promotion

and protection of the right to freedom of opinion and expression recognised that the internet is “a key means by which individuals can exercise their right to freedom of opinion and expression,” but also highlighted that existing international human rights law has put in place several standards restricting this right.⁷

DEFINING ‘COUNTER-SPEECH’

Given the overlaps between free speech, personal abuse, hate speech and violent extremism, and the tensions arising from a legal framework that identifies violations based on interpretation on a case-by-case basis, the most effective way of countering violent extremism on social media is through what Facebook calls “counter-speech.” Facebook identifies counter-speech⁸ as “crowd-sourced responses to extremist or hateful content,” based on the company’s belief that many users react to hateful posts with disagreement or argument, which retains the principle of free speech through debating on an open platform, while simultaneously tackling abusive, hateful and extremist content in an attempt to tone down the rhetoric in these spaces. The more effective instances of counter-speech appear in the comments section of the posts analysed, and revolved around appeals to recognise peaceful aspects of all religions or the defence of human rights as synonymous with national pride.

In the note on Facebook’s community standards, Marne Levine, VP of Facebook’s Global Public Policy, says, “We realize that our defense of freedom of expression should never be interpreted as license to bully, harass, abuse or threaten violence.”⁹ However, sometimes people post content that other users may consider hateful or extreme but does not violate Facebook’s policies. To counter this type of disagreeable or extremist content, Facebook has publicly stated that it believes counter-speech is not only potentially more effective but also more likely to succeed in the long run.

For researchers, the task to determine what is or is not hate speech is even more complex, especially in the Indian context. As the largest nation in South Asia, and one defined by its plurality, India is home to not only a plethora of faiths but also several other cultures and subcultures that often cut across religions. This creates many subsets of offensive or hateful speech that target both religion and cultural practices, which are often intertwined. This indicates a need to create intercultural competencies amongst various groups, promoting empathy and encouraging people to participate in positive messaging as an effective means of counter-speech. A Demos study on counter-speech on Facebook in India and elsewhere noted, “[...] religious discourse plays a central role in facilitating counter-speech discussion.”¹⁰

LEGAL APPROACHES TO HATE SPEECH IN INDIA

Even though no law in India defines what constitutes hate speech, Article 19(1) of the Constitution gives all citizens the right to freedom of speech and expression. However, these freedoms are subject to “reasonable restrictions” outlined in Article 19(2). Speech that violates, abuses or infringes in any way on “the interests of sovereignty and integrity of India, the 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 offence” is subject to censure under these restrictions.

The Indian Penal Code has several sections that deal specifically with punitive action around these reasonable restrictions. For example, Section 153(A) penalises the “promotion of enmity based on religion, race, place of birth, language.” Section 298 penalises speech that deliberately intends to wound religious sentiment. Sedition is punishable under Section 124A, and statements concerning “public mischief” under Section 505. Other laws in India that function as exceptions to the right to freedom of expression

include the Representation of The People Act 1951, the Code of Criminal Procedure 1973, the Religious Institutions (Prevention of Misuse) Act 1988, and the Scheduled Caste and Scheduled Tribe (Prevention of Atrocities) Act 1989.

The issue of the right freedom of expression online was most effectively addressed in *Pravasi Bhalai Sangathan v. Union of India* (AIR 2014 SC 1591).¹¹ Upholding the right to free speech, the Supreme Court of India asked the Law Commission if “it deems proper to define hate speech and make recommendations to the Parliament to strengthen the Election Commission to curb the menace of ‘hate speeches’ irrespective of, whenever made.” In its report on hate speech (267) released in May 2017, the Commission explained, “The standard applied for restricting Article 19(1)(a) is the highest when imposed in the interest in the security of the state.”¹² It recommended that a restriction under Article 19(2) must have a “proximate and direct” connection to a threat to public order.¹³

Although many countries have laws against hate speech, their definitions of it vary significantly. The Law Commission Report says, “The analysis of hate speech in different countries suggests that despite not having a general definition, it has been recognised as an exception to free speech by international institutions and municipal courts.”¹⁴

Efforts to combat hate speech on Facebook are particularly relevant in the Indian context. Several cases that went to court over the last five years involved individuals whose posts on Facebook had been censored or taken down for being offensive to politicians and Parliament, inciting violence, and hurting religious sentiments. These users were arrested or charged under Section 66(A) of the Information Technology Act, which aims to punish “offensive, false or threatening information” through computers and communication devices.

However, in *Shreya Singhal v. Union of India* (AIR 2015 SC 1523), the Supreme Court declared that the section “arbitrarily, excessively and disproportionately invades the right of free speech and upsets the balance

between such right and the reasonable restrictions that may be imposed on such right.”¹⁵ Thus, due its ambiguous and open-ended nature, the court declared 66(A) “unconstitutional.”

Different legal systems draw different distinctions between speech protected by ‘freedom of expression’ and ‘hate speech’. For this report, the researchers have classified speech in two ways:

1. **More lenient:** Since some definitions include abusive speech that merely insults individuals, the study marks those instances as well.
2. **More restrictive:** However, it recognises that stricter definitions of hate speech only limit speech that incites violence or discriminates based on group characteristics.

To better understand the dynamics of hate speech in a cross-jurisdictional online environment, the study draws a distinction between personal abuse and hateful speech that incites violence.

RESEARCH METHODOLOGY

Aim

This research aims to analyse the contours of hate online by examining patterns within hate speech and abusive speech on Facebook in India. Such an analysis will help identify and formulate effective responses to hate speech. While India’s judiciary draws a distinction between advocating hate and inciting violence, the general understanding of hate speech is often broader than merely the ‘incitement of violence’. If the concept of reasonable restrictions is open to interpretation in the real world, the scope for that interpretation expands significantly in the virtual world too. Therefore, to find effective strategies for counter-messaging, it is imperative to first map the composition of online hate speech. To do so, the study created a sample of hateful and abusive comments within Facebook’s India-focused communities.

Timeframes

First, two timeframes were selected to narrow the scope of the research: two one-month-long periods, approximately a year apart from each other. These timeframes were used to compare and map the trends and patterns of hate speech across the span of the year, which ensured that the analysis would not be skewed unfairly by any single event and, therefore, have broader validity across time periods and avoid a recency bias. The two time periods chosen corresponded with heightened political or social tensions, since hate speech is likely to be more prevalent during such times, both offline and online.¹⁶

The first time period was the month between 7 July 2016 and 7 August 2016 (T1). In a major counterterror operation in the Kashmir Valley, Indian security forces found and killed the Hizbul Mujahideen terrorist Burhan Wani, in the Tral area of South Kashmir on 7 July 2016. To maintain peace in the area, a curfew was put in place. However, Wani's funeral in Tral a day later became the catalyst for conflict between Kashmiri separatists and the government in the Kashmir valley, which lasted for weeks. In response to the curfew, the separatist All Parties Hurriyat Conference called for a reciprocal strike, crippling normal life for several months as schools and essential services remained closed. The standoff lasted well over a month, during which angry protesters and central-security forces clashed regularly on the streets, resulting in widespread injuries due to the use of the non-lethal pellet gun for crowd control. The protests in support of a terrorist sparked a backlash of national anger against all Kashmiris and anyone perceived as either supporting the separatists or being critical of government actions. It is important to note that the violence in Kashmir is understood widely as a longstanding political/territorial dispute and not a marker of "Islamic radicalisation."

The second time frame selected was the month between 22 June 2017 and 22 July 2017 (T2), during which three significant incidents took place. The first two happened on the same day. Seventeen-year-old Junaid Khan was lynched to death by a mob on a Mathura-bound train after an argument

over seats reportedly turned ugly. News reports suggest Junaid and his brother were mocked for their appearance (i.e., sported beards) and were accused of being “beef eaters”. The same night, a senior J&K police officer, Deputy Superintendent Ayub Pandith, was lynched to death outside Srinagar’s main Jamia Mosque on what is considered the “holiest night of the month of Ramzan,” by a mob accusing the Kashmiri Muslim police officer of plotting to kill Mirwaiz Umar Farooq, the head of the Jamia Mosque and leader of the separatist All Parties Hurriyat Conference.

The third incident was a terrorist attack on a busload of Amarnath pilgrims by Pakistan-backed Lashkar-e-Taiba terrorists in South Kashmir on 10 July 2017, which fuelled tensions in the valley and deepened existing fault-lines. The pilgrimage to the Amarnath Cave is considered sacred by the followers of Lord Shiva and is an important example of both Kashmir’s and the subcontinent’s syncretic cultural traditions (both Hindus and Muslims are custodians of the shrine). For many, the attack on the “yatis” in July 2017 was seen as the crossing of a ‘red line’. It was only the second time they had been attacked since the insurgency began in Kashmir 28 years ago.

Pages and Categories

A list of public pages was curated to represent Facebook’s India presence. The researchers chose 400 pages—50 pages across eight categories—that include a variety of stakeholders, opinions and communities:

1. **National Mainstream Media** - This included print, broadcast and web-based news outlets and were chosen based on overall popularity measured by views, readership and Facebook likes.
2. **Political Organisations and Personalities** - Every major political party in India’s Lok Sabha that met the specific criteria was included, in addition to prominent politicians from across the political spectrum.

The major student wings of political parties, too, were selected. Prominent politicians in Jammu and Kashmir are excluded here; they are included in the Kashmir-focused category.

3. **Alternative News and Opinion Pages** - Pages in this category focused on discussing and presenting current affairs outside the mainstream media framework, and the content displayed somewhat clearer partisanship. This was redressed by including different shades of opinion to capture all sides of the conversation.
4. **Religious Organisations and Personalities** - Major religious organisations from all of India's faiths were represented, including temples, mosques, churches and gurudwaras. In addition, prominent religious leaders with active Facebook pages were chosen.
5. **Satire and Humour** - Pages that commented to some extent on current affairs were chosen, so as not to stray from the general political discourse. Individual comedians are also part of this category.
6. **Community Organisations, Causes, Charities and NGOs** - This category includes more niche communities—both online and offline—that focus on particular issues. These pages are generally smaller but generate specific discourse that is valuable and were chosen to be representative.
7. **Prominent Personalities** - Several important Indian commentators are not religious leaders, politicians or comedians, and are included in this category. These individuals were selected based on their political statements in the past or frequent commentary on current affairs. They include sportspeople, journalists, actors and other members of civil society.
8. **Kashmir Focused** - The incidents that motivated the selection of the time periods focused on Kashmir, and it was important to include a more detailed coverage of the area in the study. Pages in this category include Kashmir-based media outlets, commentary pages and prominent Kashmiri politicians and separatist leaders.

For the study, pages that had at least 20 posts in each time frame were chosen, to provide sufficient data to analyse. However, since the majority of Facebook pages post less frequently than this, the requirement was eased in some cases. Therefore, 47 of the 400 pages have between 10 and 20 posts in each month. Due to limitations in analysing other scripts using the search method, the study only selected pages where most posts and comments are in Roman script.

There were three exceptions to this methodology that did not meet these criteria. However, they were included since they form a crucial part of the Indian discourse. These were:

1. **Barkha Dutt:** A prominent Indian journalist (who did not have 10 posts in each time frame, but had several thousand comments on the few that she did)
2. **Republic:** A leading news television channel that frequently tops viewership ratings in the English news market (that did not exist in T1)
3. **Syed Ali Shah Geelani:** One of the most prominent Kashmiri separatist leaders (whose current page did not exist in T1, as it replaced an older, now-deleted page)

Data Collection and Further Sampling

Facebook's Graph API was used to scrape every post in T1 and T2 from the 400 pages, with a self-modified version of a freely available Python script¹⁷ to restrict the posts scraped by date.

The number of posts gathered varied widely: one mainstream media source had over 14,000 posts in one month, whereas several pages had between 10 and 20. To ensure that the analysis was not completely overpowered by larger pages, as well as to maintain the representative value of the dataset, researchers used the statistical programming language R to

sample 20 posts from each page. For pages with fewer than 20 posts, every post was used. After this process had been completed, the study’s sample size had 15,341 posts in all, with 7,712 in T1, and 7,629 in T2.

Preliminary inspection of India-related Facebook pages indicated that significant amounts of hate speech and counter-speech were present in the comments on posts, and that the posts themselves had less objectionable content. It is perhaps the case that since the posts were higher visibility, they attracted more attention from vigilant community moderators, leading to hateful posts being removed.

Another freely available Python script by the same author was modified to scrape the comments from each of the approximately 15,000 posts in the dataset. The number of comments were highly variable as well, with some posts attracting thousands of comments and others getting none. The study did not, however, restrict the number of comments as it did with posts, since that would arbitrarily remove comments that were replies to others in the dataset. Since several conversations and debates on Facebook take place using the reply function, removing comments arbitrarily would harm the ability to understand the debate that is crucial to countering hate speech.

Several comments were posted outside of the timeframes of interest, because Facebook permits users to comment on posts well after they have been posted. These were excluded from the analysis. The numbers of comments scraped are given below:

	T1: 7.7.16–7.8.16	T2: 22.6.17–22.7.17
Posts	7,712	7,629
Comments	617,954	767,924
Comments in Time Frame	523,395	674,077
Average Comments in Time Frame per Post	67.8	88.4

Finding Hate Speech and Counter-Speech

The next step was to extract comments likely to contain hate speech. The researchers selected the comments that included keywords appearing frequently in hateful and abusive replies. They conducted a dipstick analysis on 10 percent of the posts, reading both the posts and the comments on those posts to compile the list of search terms.

A total of 220 search terms were compiled, including alternate spellings for terms in Hindi and Urdu that have no standard English spellings. While several of these terms were inherently profane, some were innocuous yet used frequently in hate speech.

These search terms were used to filter the dataset and collect a final set of comments containing them.

	T1: 7.7.16–7.8.16	T2: 22.6.17–22.7.17
Comments with Search Terms	6,314	8,376

Content Analysis

Using the method of Content Analysis (CA), the team sought to understand patterns in hate speech and counter-speech in both T1 and T2. CA can be defined as the “systematic, objective, quantitative analysis of message characteristics.”¹⁸ It helps classify and better manage the data through a well-defined process that allows researchers to draw valid inferences. In this study, CA helped deduce credible patterns from the representative samples while eliminating presenter or reader-interpretation bias.

CA is made effective by using Inter Coder Reliability (ICR). ICR refers to the extent of agreement between a minimum of two researchers (Neuendorf 2002, p. 144) with the goal that all researchers code the same content with the same values. The “extent of agreement” is considered to be

an indicator of the data's reliability. Two researchers read the same sample of 150 comments to calculate ICR using the interrogation framework below. Calculating ICR as $\text{agreements}/(\text{agreements} + \text{disagreements})$ led to an ICR of 86 percent (126/150). For content analysis, a coefficient of .80 or greater is acceptable in most situations.¹⁹

Using R, the research team sampled randomly, and then coded, 25 percent of the comments in each time frame, leading to 1,579 comments in T1 and 2,094 comments in T2.

The interrogation framework to analyse these comments is given below. The study adopted the form of 12 Yes/No questions. The first two questions have mutually exclusive responses, i.e. no comment could be coded both as hate speech and personal abuse, as it was assumed that any hate speech would contain abusive characteristics. The other 10 questions were only answered for comments that were hateful or personally abusive.

1. Is this hate speech?
2. Is this personal abuse?
3. Does it mention a particular religion?
4. Does it talk about religio-cultural practices?
 - a. Food
 - b. Dress
 - c. Practices
5. Does the comment question your nationality and your commitment to your nation?
6. Does the comment use profanity?
7. Does it incite direct bodily harm?
8. Does it call for mass violence against a particular community/group?
9. Does it mention or threaten gender-based violence?
10. Does it mention LGBT+ people?
11. Does it mention a political party?
12. Does it advocate/incite theft, vandalism or financial harm?

Several questions were framed keeping in mind Facebook's own standards, particularly questions 7, 8, 9, 10 and 12. Questions 1 and 2 distinguish between hate speech and personal abuse. Earlier in the paper, the theoretical framework discusses the manifold interpretations of hate speech from legal and corporate points of view. Facebook separates bullying and harassment from hate speech in its Community Standards. The preliminary analyses indicated that there were significant differences between the tone and content of comments and posts that were hateful and those that were abusive.

The criteria that the researchers used in answering these questions are as follows:

- Any explicit threat, incitation or advocacy of violence against a person or group of people was marked as hate speech.
- Comments that attacked groups or individuals based on their community characteristics (including religion, gender, caste, nationality and sexuality) were marked as hate speech.
- Comments that attacked groups or individuals for other reasons, including disagreement and heated arguments, were marked as personal abuse. This included profanity targeted against individuals.

Questions 3 and 4 are specific to the study of hate speech in India. Under Indian law, reasonable restrictions are placed on freedom of speech when it pertains to attacks on specific religions or religious/cultural practices that discriminate between communities: Hindus and Muslims, for the purpose of this study. A nationwide debate over the banning of cow slaughter and eating beef has sharpened the faultlines of religion and culture in India, and Junaid's lynching and the surrounding commentary is indicative of "religio-cultural hate" that discriminates against or attacks an individual or group on the basis of religion, dress and other perceived or real cultural/food habits. There were several comments that mentioned practices but not the religion explicitly.

Question 5 was included based on the time frames of the study. There has been significant anecdotal evidence to indicate that Kashmiris, or those

accused of speaking on behalf of Kashmiris, have been popularly called “anti-national” or told to “go to Pakistan,” a narrative that is specific to the Indian context.

Question 10 on gender and sexual identity, while a significant category of Facebook’s own community standards, takes on added relevance in the Indian context. Under Section 377 of the Indian Penal Code, homosexuality is a criminal, punishable offence in India.

LIMITATIONS OF THE STUDY

- Any study that aims to encourage counter-speech must be able to identify content in the several vernacular languages used in India. The analysis in this study was limited because it only included pages that post in English and some instances of Hindi or Urdu comments, only if written in the Roman script. A deeper, more thorough understanding of hate speech on Facebook must be conducted in regional languages too.
- A second significant obstacle in trying to understand the spread of hate speech and the efficacy of counter-speech on Facebook is that the study is limited to public pages that can be accessed by anyone and are open to comments and posts from any of its two billion plus users. Thus, while this study has unearthed some abhorrent, vitriolic and violent messaging, those on private/closed pages of individuals or communities could be even stronger.
- In times of a security or a law-and-order problem, the Indian government often responds by shutting down the internet. Mobile internet connectivity was deactivated in Kashmir for the entire duration of T1, between 7 July and 7 August 2016. Additionally, several posts, especially about Kashmir, have already been blocked or removed by Facebook.

- While hate speech and discrimination based on caste is an entirely different and equally prominent category of abuse, it was left out of the scope of this study as the comments during T1 and T2 did not refer to any significant caste-based abuse. However, incidents of violence triggered by caste conflict or national debates on reservations for Scheduled Castes and Scheduled Tribes might result in heightened commentary along these sociocultural fault lines. This would require a separate study, as would an analysis of gender-based hate.

RESULTS/FINDINGS

Since several of the search terms can be used outside of hateful contexts, a significant proportion of the comments analysed were not coded as hateful or abusive. However, sufficient numbers of comments were obtained, which are illustrative for the purposes of this study, as indicated below.

	T1	T2	T1	T2	T1 + T2
	(25% sample)	(25% sample)	(Extrapolated to full dataset)	(Extrapolated to full dataset)	(Extrapolated to full dataset)
Comments Analysed	1,579	2,094	6,314	8,376	14,690
Hate Speech	232	267	928	1,068	1,996
Personal Abuse	257	318	1,028	1,272	2,300

Hate Speech

The breakdown of the results for hate speech comments can be identified in two trends:

1. Over the span of one year, religion has emerged as the more explicit

basis for hate: a jump from 19 to 30 percent. The upswing of comments on food also relate to this, with the beef-ban discourse polarising commentary along religio-cultural lines.

2. Responses to hate speech around violence in Kashmir have been broadly framed in a nationalistic “India vs Pakistan” context.

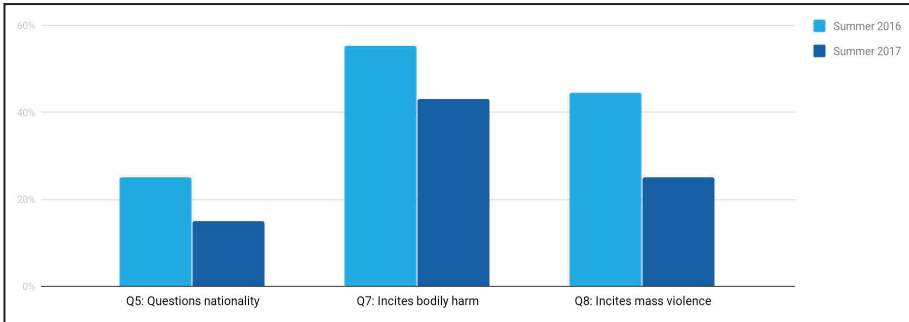
	T1	T1	T2	T2
		(% of total)		(% of total)
Total Hate Speech	232	100%	267	100%
Q3: Mentions religion	44	19.0%	80	30.0%
Q4: Mentions religio-cultural practices	31	13.4%	50	18.7%
Q4a: Food	2	0.9%	27	10.1%
Q4b: Dress	1	0.4%	4	1.5%
Q4c: Other practices	28	12.1%	40	15.0%
Q5: Questions nationality	59	25.4%	41	15.4%
Q6: Uses profanity	72	31.0%	100	37.5%
Q7: Incites bodily harm	128	55.2%	115	43.1%
Q8: Incites mass violence	103	44.4%	67	25.1%
Q9: Mentions gender violence	17	7.3%	13	4.9%
Q10: Mentions LGBT+ people	2	0.9%	1	0.4%
Q11: Mentions political party/politician	14	6.0%	25	9.4%
Q12: Advocates financial harm	9	3.9%	10	3.7%

A significant proportion of comments in both T1 and T2 incite bodily harm; it is the largest classification. Of them, more than half advocate violence against a group of people. Across time frames and the specific incidents, two principles are apparent:

- Hate speech on Facebook is reactive: Those who interact with posts do so with an eye on current affairs and respond to the news cycle.

- Counter-messaging must be equally nimble and respond to the narratives of hate that surface in the context of the news in question.

Percentage of Hate Speech comments in each time frame that is related to nationalism and/or violence

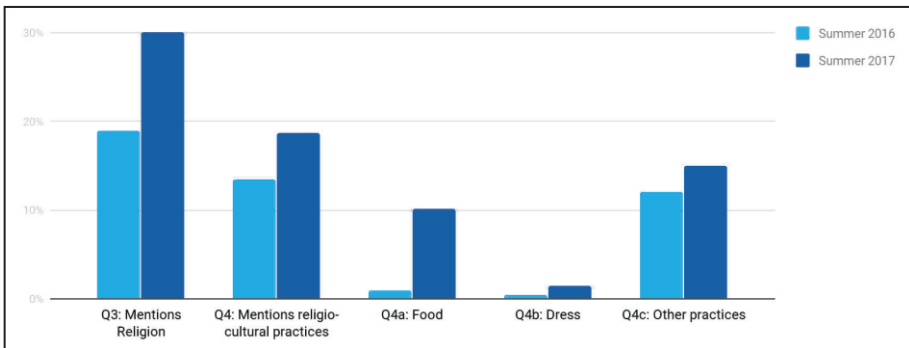


The analysis also identifies some key differences across time frames that help explain the patterns of hate speech on India-oriented Facebook pages.

- Mentions of religion and religio-cultural practices are significantly higher in T2. This likely results from the increased references to the religious character of violent incidents that took place in it. Specifically, the more than tenfold rise in mentions of religio-cultural food practices coincides with the allegations that the lynched teenager, Junaid, ate beef.
- Questions surrounding nationality and patriotism declined from T1 to T2, e.g. fewer instances of individuals or groups being called “traitors.” The nature of the uprisings in Kashmir after Burhan Wani’s death likely spurred this, since there were heavy protests against the Indian government. Given the protests, there was a preoccupation with measuring commitment to the nation during T1, and more hate speech was targeted at those believed to have betrayed the country. This is reflective of the Kashmir dispute as the key focus of a much larger India vs Pakistan narrative.
- The incitement of bodily harm and the advocacy of mass violence

were both lower during T2. Such speech is particularly worrisome and is clearly illegal in most jurisdictions, irrespective of interpretations around violations of the freedom of expression. In the dataset analysed here, this violence was directed primarily at protesters in the Kashmir Valley.

Percentage of Hate Speech comments in each time frame that is related to religio-cultural practices



The study further analysed posts that attracted the most calls for mass violence: five posts had over five comments of this nature (four in T1 and one in T2). Each one of these posts discussed protests in Kashmir, and four had an explicitly biased, but not necessarily hateful, presentation of the issue.

The post that attracted the most calls for mass violence (13) posed an open question to its readership asking “what our [Indian] soldiers are supposed to do” in response to the protests. This goaded several users to post comments calling to “kill them,” and “shoot them.” One post had a photograph of stone-pelting protesters, and some with the Pakistani flag. Another presented a video of army soldiers being harassed by locals in Kashmir during the protests, again provoking comments advocating harsh violence against Kashmiri locals.

This provides another valuable insight: the presentation of a perceived enemy leads to violent threats against them, more so than simply presenting news. During T2, there was no visible “enemy,” and so the impetus to threaten was reduced.



Personal Abuse

1. Across both time frames, most personal abuse targeted the individual's patriotism.
2. Profanity was widespread in personal comments.

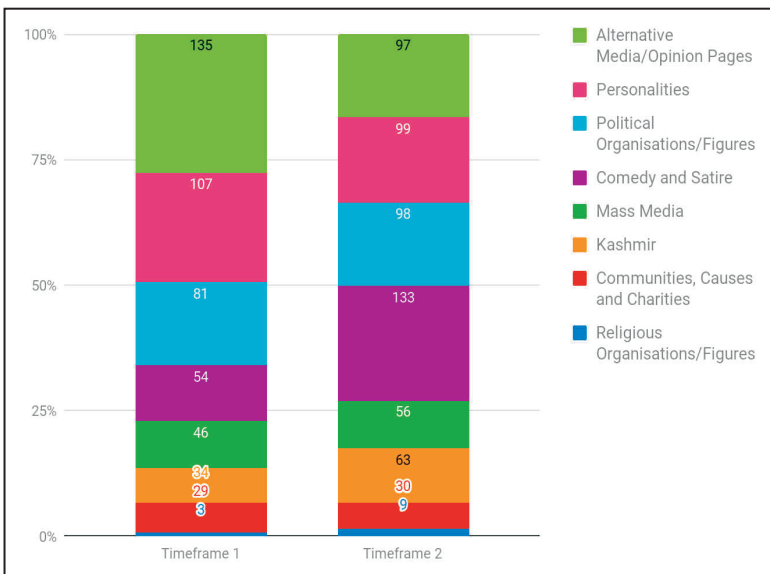
	T1	T1	T2	T2
		(% of total)		(% of total)
Total Personal Abuse	257	100.0%	318	100.0%
Q3: Mentions religion	9	3.5%	6	1.9%
Q4: Mentions religio-cultural practices	3	1.2%	18	5.7%
Q4a: Food	0	0.0%	7	2.2%
Q4b: Dress	2	0.8%	1	0.3%

Q4c: Other practices	2	0.8%	15	4.7%
Q5: Questions nationality	42	16.3%	50	15.7%
Q6: Uses profanity	180	70.0%	251	78.9%
Q9: Mentions gender violence	2	0.8%	2	0.6%
Q10: Mentions LGBT+ people	2	0.8%	3	0.9%
Q11: Mentions political party/politician	30	11.7%	20	6.3%
Q12: Advocates financial harm	3	1.2%	1	0.3%

Questions 7 and 8 are excluded since they had no responses (all violent threats are automatically classified as hate speech.) What is striking about this analysis is the high proportion of comments that use profanity, far more than that in hate-speech comments. Researchers must thus be careful when distinguishing between hate speech and personal abuse. Considering profanity inherently “hateful,” even if directed against an individual, dilutes actual analysis of the issue.

Page Characteristics

Hate Speech and Personal Abuse by Page Category



The graph above displays the number of abusive and hateful comments by the category of the page on which they appeared. The relative lack of commentary—both hate speech and counter-speech—during T1 in the dataset of Kashmir-focused pages is explained by the shutdown of mobile internet services in Kashmir.

Religious pages and single-issue communities and causes host a very low proportion of both types of comments. Instead, broader pages serve as locations for this kind of discourse, with alternative media sources, comedy pages and prominent personalities taking lead. This seems due in part to the potential for disagreement on these pages. Pages with a narrower focus tend to only attract those who already agree with the views the page puts across, and even when there is disagreement on those pages, it is more civil. Pages with a broader audience, on the other hand, have ideologically and culturally diverse commenters, which leads to the potential for more violent, and therefore hateful, arguments.

Effective counter-speech must take place on pages that serve as a common meeting ground, rather than stay restricted to niche pages that become echo chambers for like-minded individuals and institutions.

COUNTER-SPEECH

While the commenter above clearly showed dislike of the hateful content, she did not discuss the subject ideas, and so the comment was **ineffective**.

On the other hand, the last commenter here (in blue) specifically responded to the hateful commenter (in red). They addressed the religio-cultural nature of the hatred (see "Allah-u-Akbar"), and promoted a message of tolerance and peace. His message was well received, and was thus **effective** counter-speech

As mentioned earlier, Facebook identifies counter-speech as crowd-sourced responses to extremist or hateful content, based on the company’s belief that open interaction allows for debate and argument, and attempts to tone down the pitch.

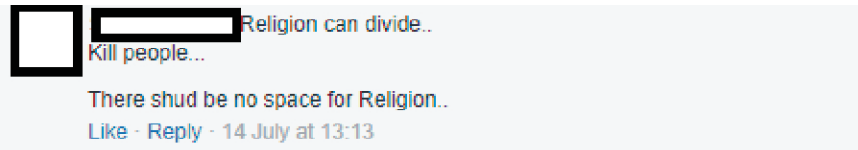
LYNCHING IS HORRIFIC CRIME...ITS A GROWING MENACE IN ALL COUNTRIES ...WHERE POLICE LAWS WEEK...WE SHOULD CONCENTRATE ON MAKING STRICT POLICE LAWS..INSTEAD OF RELIGION BLAH BLAH...IN ANY LYNCHING CASE
 Like · Reply · 28 June at 16:03

Kashmiri pundits being denied their place of birth is a crime. No one has the rights to deny their right to live in Kashmir. This year kashmiri muslims opened a temple, cleaned it n observed shiv ratri...hope u remember the news. Hundred of them held placards calling for the return of the pundits could be seen. Don't u think the govt should have taken advantage of this conducive situation to call back the pundits?! They won't...cos if this issue is solved then bhakts wont have this reason to justify their murders. N ppl won't have a reason to comment here.
 Like · Reply · 8 July at 10:55

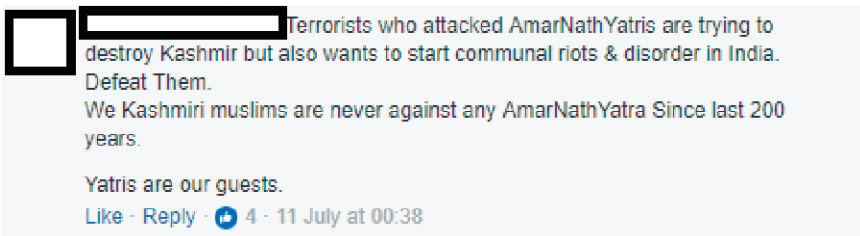
Jihad mazloom ki maddad kay liya kiya jata hay jesay Kashmir may Muslim india ka dalits etc lekin kisi ko na haq nahi mara ja sakta yeh mana hay
 Like · Reply · 22 July 2016 at 09:44




Counter-speech is not restricted to a single narrative. While some of the comments in the study were standalone, others were replies to comments. Some of the counter-speech comments appeal to religion, stating that religions (either a specific religion or all religions inherently) are meant to be peaceful and protect innocent people. Lynching in the name of cow protection and terrorism in the name of Jihad were both denounced as not being a part of Hindu or Islamic religious doctrines respectively.

A narrative that emerged and needs to be encouraged was one that appealed to a sense of common decency and humanity, condemning the killing of all innocents. Such comments played down the role of religion, caste, communities and ethnicity as divisive. Some comments also condemned the tendency to stereotype people from religious groups, and criticised institutions such as the media and political parties, accusing both of dividing people for vested interests.


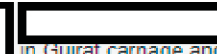



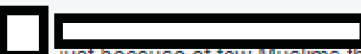

A lot of counter-speech took place around the protests in Kashmir. Commentary ranged from appeals for peace and calm, to advocating for a change in the government's approach, emphasising the need for a more humane method. Others condemned acts of terrorism and the killing of innocent people. Some of the counter-speech also called upon people to avoid, and/or be wary of, extremist agendas.




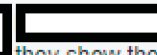

  Kashmir is all. Kashmir is Love.
In Sha Allah Kashmir will blossom all over again and this time hopefully, let us all pray that in future it never witnesses any turmoil any more.
Let this be the end of darkness and beginning of the new dawn.
May now the light, love and peace caress this nation.
Our nation 'KASHMIR '
Amen.
Like · Reply ·  4 · 1 August 2016 at 07:49



The study came across examples of users who tried to bridge discrimination and explain the distinction between orthodoxy and radicalisation in the context of Muslims vs Islamic extremism, as well as in the context of lynch mobs in the name of cow protection vs Hinduism. Some directly took on religio-cultural hate speech, challenging notions of discrimination based on food or dress habits.

  Cowardly attack on innocent people as it has happened in Gujrat carnage and various lynching in different regions of India.Cowards everywhere attack on unnamed civilians.Violence must be condemned at every level.
Like · Reply · 12 July at 00:53

  Muslims are not terrorists brother it's just because of few Muslims the name of the entire community is getting spoilt please learn to respect the religion.
Like · Reply ·  8 · July 28, 2016 at 12:40am

Overall, most counter-speech comments called for peace and attempted to end the cycle of hatred.

  No matter where u go haters will be always there..and they show their class with language they use 😊
Like · Reply ·  4 · August 2, 2016 at 11:21pm

  Hindu Muslim Bhai bhai
Like · Reply · 27 June at 04:54

LESSONS FOR INDIA

1. Hate speech is reactive and follows current events closely. During T1, the narratives centred around the concept of ‘nationhood’, in the backdrop of protests against the Indian government. During T2, the discourse centred on religio-cultural factors, made relevant due to the religious characteristics of the incidents at the time.
2. Hate speech takes place primarily on “melting pot” pages where people with different beliefs clash.
3. When you present partisan audiences with provocative images or videos of the “enemy,” incidents of violent speech increase.
4. It is important to differentiate between personal abuse and hate speech. Much hate speech does not contain profanity, so that alone cannot serve as a measure of acceptability.
5. Violent speech on gender was less prevalent during T1, but every time the LGBT+ community was mentioned, the reference was derisive.

LESSONS FOR COUNTER-SPEECH

1. Counter-speech and counter-narratives must focus on current affairs and directly address present concerns. Otherwise, they risk being irrelevant.
2. Counter-speech must concentrate on larger, more diverse communities instead of seeking out niche, smaller pages.
3. Narratives that can smooth tensions between religious communities are more successful than those that address political differences.
4. Calls to reduce the pitch of rhetoric, without necessarily offering a counterpoint, are prevalent but not effective.
5. Counter-speech is effective when it addresses ideas, not specific content.

6. Administrators of mainstream pages with large followings must be encouraged to identify, highlight and promote counter-speech.
7. Counter-speech tends to take place more in comments than in posts. Therefore, through Facebook's own messaging and content, users can be encouraged to become a part of an ecosystem that promotes dialogue and community.

(The authors thank Raghav Bikhchandani, Ashini Jagtiani, and Radhika Jhalani for their research assistance.)

ABOUT THE AUTHORS

Maya Mirchandani is Senior Fellow at Observer Research Foundation, where she heads the Countering Violent Extremism project. She is a broadcast journalist with over two decades of reporting with NDTV.

Ojasvi Goel is a student of economics at the London School of Economics and Political Science. He worked on this report while interning at ORF.

Dhananjay Sahai is a student of Law at Delhi University's Campus Law Centre. He worked on this report as a Research Assistant at ORF.

ENDNOTES

1. Levine, Marne. "Controversial, Harmful and Hateful Speech on Facebook". A statement on Facebook from the company's VP, Global Public Policy. <https://www.facebook.com/notes/facebook-safety/controversial-harmful-and-hateful-speech-on-facebook/574430655911054/>.
2. Germany recently passed the Network Enforcement Act, popularly known as the "Facebook Law," which requires social-media companies such as Facebook and Twitter operating in Germany to block or delete any kind of hate speech, and racist or slanderous comments or posts within 24 hours of being reported by users.
3. Allan, Richard. "Hard Questions: Hate Speech," Facebook Newsroom, 27 June 2017. <https://newsroom.fb.com/news/2017/06/hard-questions-hate-speech/>.
4. Ibid.
5. Whitman, James Q., "Enforcing Civility and Respect: Three Societies" (2000). Faculty Scholarship Series. 646. http://digitalcommons.law.yale.edu/fss_papers/646
6. Gagliardone, Iginio, Gal, Danit, Alves, Thiago, and Martinez, Gabriela. "Countering Online Hate Speech", UNESCO Series on Internet Freedom, 2015. <http://unesdoc.unesco.org/images/0023/002332/233231e.pdf>.
7. La Rue, Frank. "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression". Human Rights Council, 17 Session. 16 May 2011. <https://documents-dds-ny.un.org/doc/UNDOC/GEN/G13/133/03/PDF/G1313303.pdf?OpenElement>
8. Bartlett, Jamie and Krasodomski-Jones, Alex. "Counter-speech on Facebook," Demos. September 2016. <https://www.demos.co.uk/wp-content/uploads/2016/09/Counter-Speech-on-facebook-report.pdf>.
9. Levine, Marne. "Controversial, Harmful and Hateful Speech on Facebook". A statement on Facebook from the company's VP, Global Public Policy. <https://www.facebook.com/notes/facebook-safety/controversial-harmful-and-hateful-speech-on-facebook/574430655911054/>.
10. Bartlett, Jamie and Krasodomski-Jones, Alex. "Counter-speech on Facebook," Demos. September 2016. <https://www.demos.co.uk/wp-content/uploads/2016/09/Counter-Speech-on-facebook-report.pdf>.
11. Pravasi Bhalai Sangathan vs U.O.I. & Ors on 12 March, 2014. <https://indiankanoon.org/doc/194770087/>.
12. "Hate Speech". Law Commission of India. Report No. 267. March 2017. <http://lawcommissionofindia.nic.in/reports/Report267.pdf>.
13. Paragraph 3.6 AIR 2014 SC 1591.
14. Paragraph 4.8 AIR 2014 SC 1591.
15. Shreya Singhal vs U.O.I on 24 March, 2015. <https://indiankanoon.org/doc/110813550/>.
16. Samir Saran, "ISLAM – A Threat to the West: How has the British Press portrayed Islam as a threat to the West in the aftermath of the '7/7' London bombings?" Unpublished essay, Observer Research Foundation.
17. Woolf, Max, Facebook Page Post Scraper, (2017), GitHub repository, <https://github.com/minimaxir/facebook-page-post-scraper>.
18. Neuendorf, Kimberly A, The Content Analysis Guidebook. Sage Publications, 2002.
19. Joyce, Mary. "Picking the Best Intercoder Reliability Statistic for Your Digital Activism Content Analysis," <http://digital-activism.org/2013/05/picking-the-best-intercoder-reliability-statistic-for-your-digital-activism-content-analysis/>.

Observer Research Foundation (ORF) is a public policy think-tank that aims to influence formulation of policies for building a strong and prosperous India. ORF pursues these goals by providing informed and productive inputs, in-depth research, and stimulating discussions. The Foundation is supported in its mission by a cross-section of India's leading public figures, including academic and business leaders.



20, Rouse Avenue Institutional Area, New Delhi - 110 002, INDIA
Ph. : +91-11-43520020, 30220020. Fax : +91-11-43520003, 23210773
E-mail: contactus@orfonline.org
Website: www.orfonline.org