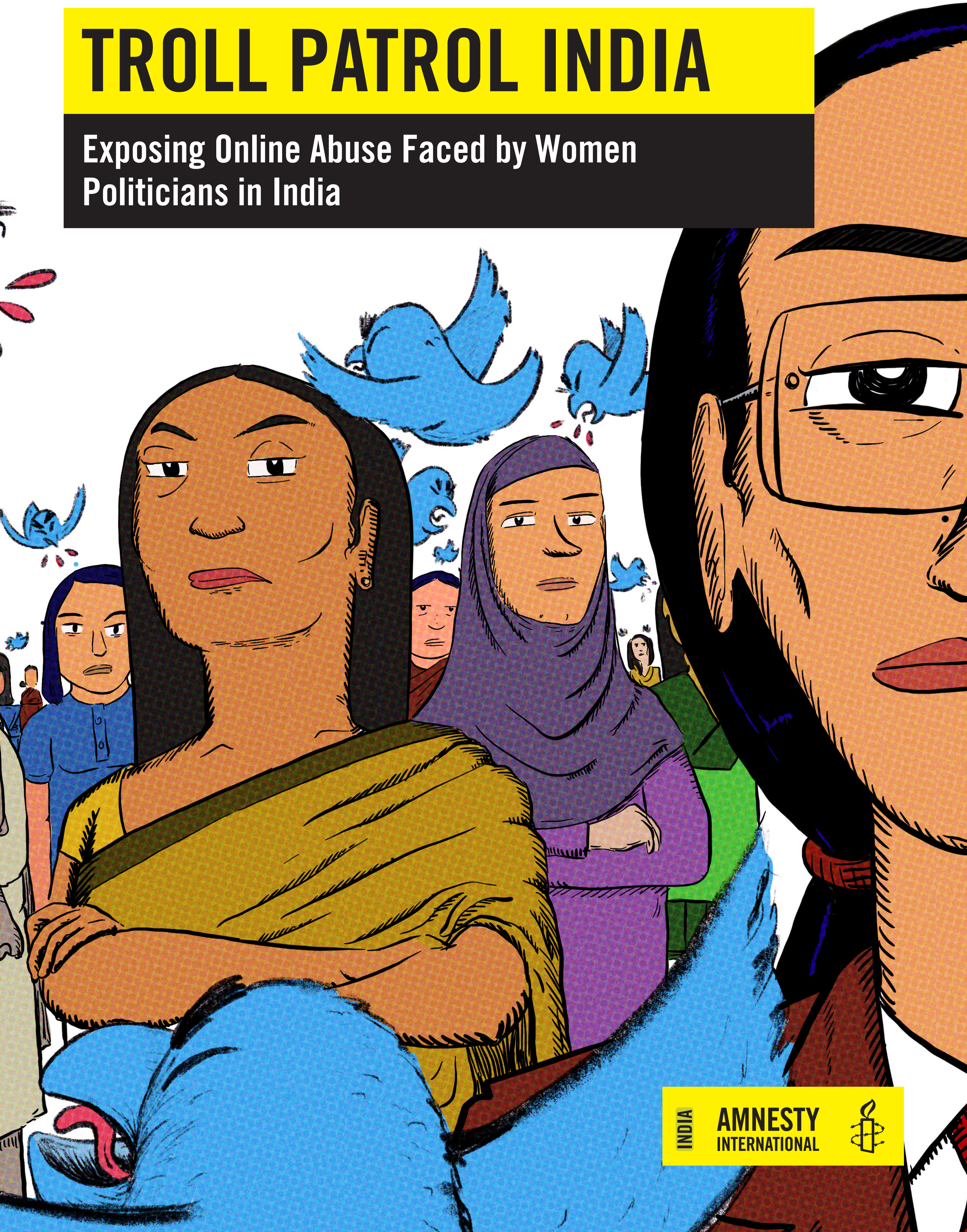
TROLL PATROL INDIA

Exposing Online Abuse Faced by Women Politicians in India



INDIA AMNESTY INTERNATIONAL

Amnesty International India is part of the Amnesty International global human rights movement. Amnesty International India seeks to protect and promote the human rights of everyone in India. Our vision is for every person in India to enjoy all the rights enshrined in the Universal Declaration of Human Rights, other international human rights standards and the Constitution of India. We are independent of any government, political ideology, economic interest or religion, and are funded mainly by contributions from individual supporters.

INDIA AMNESTY INTERNATIONAL

First published in 2020 by
Indians For Amnesty International Trust
#235, Ground Floor, 13th Cross, Indira Nagar, 2nd Stage,
Bengaluru – 560038, Karnataka, India

Original language: English

Illustrations: Bakarmax

TROLL PATROL INDIA

Exposing Online Abuse Faced by Women Politicians in India

CONTENTS

Executive Summary

इसे बदचलन बोलते हैं I चाल चरित्र खराब लोगो की यही लक्षण
Rakhel
ko bhi kesh ke hi padega
अब तुझे अपनी जात और औकात का आभास जनता करा देगी
Are gadari ke had hoti ha ek Randi ka y iman hota ha
Chappal padna um jaise h ko
Suar ka bal hai iski g*** main
Badsurat chehra
Rakh
Is Kutiya ko bhi Gauri lankesh ke pas bejna hi padega
थर्ड क्लास औरत तू अपना पास्ट भूल गई तू कैसे राजनीति में ऊपर आयी है ये सबको पता है
Badsurat chehra
arna kamre mei le doonga

India is one of the largest and fastest growing audience markets globally for Twitter, a social media platform.[1] Touted as a 'safe place for free expression', Twitter was envisioned to be a space where marginalised populations, including women, Dalits and religious minorities, would have an equal opportunity to make their voices heard. Over the years, the social media platform has evolved into an indispensable tool for political engagement, campaigning and activism, but has the vision translated into reality? Many women do not believe so. Every day, women on Twitter face a barrage of abuse: from racist and sexist attacks to rape and death threats.

Using crowd-sourced research and data science, Indians for Amnesty International Trust,[2] in collaboration with Amnesty International – International Secretariat (AI-IS) measured the scale and nature of online abuse faced by women politicians in India during the 2019 General Elections of India. The study found that abuse experienced by Indian women politicians was high, suggesting that Twitter is failing in its responsibility to respect women's rights online. The study supports the notion that for many women, the social media platform has turned into a 'battlefield'.[3]

We studied tweets mentioning 95 Indian women politicians in the three-month period of March – May 2019, in the lead-up to, during and shortly after the 2019 General Elections in India. Of the total volume of 7 million tweets mentioning these politicians, we sampled 114,716 for the purpose of analysis through our Troll Patrol India project.[4]

Engaging 1,912 volunteers, known as 'Decoders' from 82 countries, the tweets were analysed to create a labelled set of 'problematic' or 'abusive' content. The Decoders were shown a tweet with username obscured, mentioning one of the women in our study. They were, then asked simple questions about whether the tweet was problematic or abusive, and if so, whether they revealed sexist or misogynistic, religious, casteist, racist or homophobic abuse, or physical or sexual threats. Each tweet was analysed by multiple people. The Decoders were given a tutorial, definitions and examples of 'problematic' and 'abusive' content, as well as an online forum where they could discuss the tweets with each other and with our researchers. The labelling of the Decoders was analysed between July 2019 - November 2019.[5]

PROBLEMATIC CONTENT

Tweets that contain hurtful or hostile content, especially if repeated to an individual on multiple occasions, but do not necessarily meet the threshold of abuse. Problematic tweets can reinforce negative or harmful stereotypes against a group of individuals (e.g. negative stereotypes about a race or people who follow a certain religion).

We believe that such tweets may still have the effect of silencing an individual or groups of individuals. However, we do acknowledge that problematic tweets may be protected expression and would not necessarily be subject to removal from the platform.

ABUSIVE CONTENT

Tweets that promote violence against or threaten people based on their race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. Examples include physical or sexual threats, wishes for the physical harm or death, reference to violent events, behaviour that incites fear or repeated slurs, epithets, racist and sexist tropes, or other content that degrades someone.

> **"When you are on social media, you face trolls, threats, abuses and challenges 100% of the time. Their purpose is to silence you. It makes you want to cry. They talk about your personal life, your looks, and your family."**
>
> – *Alka Lamba, Member, Indian National Congress*

[1] Colin Crowell, Setting the record straight on Twitter India and impartiality, 8 Feb. 2019, Twitter, https://blog.twitter.com/en_in/topics/events/2019/impartiality.html

[2] Hereinafter referred as 'Amnesty International India'

[3] Interview with Shazia Ilmi, Member, Bharatiya Janata Party, on 25 November 2019 in New Delhi, India

[4] For purpose of study, the term 'women politicians' includes party members who may or may not have held elected office

[5] The crowd-sourced research mobilised as many as 1,912 Decoders from 82 countries, with 57.3% (1,095 Decoders) hailing from India across 26 states. About 1,750+ hours was contributed cumulatively by the Decoders. The research analysed tweets in 9 languages, including Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Tamil and Telugu.

WHAT DID WE FIND?

1. **1 in every 7 tweets that mentioned women politicians in India was 'problematic' or 'abusive'**

   13.8% of the tweets that mentioned 95 women politicians in the study were problematic (10.5%) or abusive (3.3%). This amounted to 1 million problematic or abusive mentions of the 95 women between March and May 2019, or over 10,000 problematic or abusive tweets every day across all women in the sample.

2. **Indian women politicians experienced substantially higher abuse than their UK and USA counterparts**

   A similar study conducted by Amnesty International in the UK and USA in 2018 to measure the online abuse faced by 323 women politicians found that 7.1% tweets mentioning politicians were 'problematic' or 'abusive'.[6] Whereas this study, using a similar methodology, but focusing on a shorter period during elections found that Indian women politicians experienced 13.8% problematic or abusive tweets, which is substantially higher.

3. **Women politicians prominent on Twitter are targeted more**

   It has long been assumed that being more visible on social media as a woman, can lead to more abuse. This study confirms that there is a clear correlation between the number of mentions and the proportion of abusive content received by women.

   We tested this by grouping the politicians by mentions. When we split the group into the top 10 most-mentioned politicians and others, we found that the top 10 had a mean of 14.8% problematic or abusive content versus 10.8% for the others. This meant that while the top 10 received 74.1% of all mentions, they received 79.9% of problematic or abusive mentions.

4. **1 in every 5 problematic or abusive tweets was sexist or misogynistic**

   Decoders were directed to label the problematic or abusive content as containing sexism and/or misogyny, ethnic or religious slur, racism, casteism, homophobia or transphobia, sexual threats, physical threats or 'other'. 'Other' was used when tweets containing problematic or abusive content did not fit in the categories provided. Notably, the most common selection was the 'other' category (74.1%). This indicated that most problematic or abusive content did not fit neatly into any of the suggested categories. Nevertheless, of all the tweets that were labelled as problematic or abusive, over 1 in 5 answers (19.9%) showed sexism or misogyny.

   A deeper look at the tweets that demonstrated sexism or misogyny showed that sexism was experienced by women independent of political ideology or affiliation, religion, caste, race, age, marital status, and election outcome.

5. **Muslim women received 94.1% more ethnic or religious slurs than women from other religions**

   Women politicians who are or perceived as Muslims received significantly more abuse when compared to women from other religions. They received 55.5% more problematic or abusive content. 26.4% of the problematic or abusive content experienced by them contained ethnic/ religious slurs, nearly double the proportion for women who are or perceived as Hindus (13.7%).

6. **Women politicians belonging to marginalised castes received 59% more caste-based abuse compared to women from other castes**

   Women from marginalised caste received 59% more caste based abuse than women from general castes. In cases, where problematic or abusive content was identified, women belonging to marginalised castes, such as Scheduled Castes, Scheduled Tribes and Other Backward Classes (8.6%) received more caste-based slurs than those belonging to general (5.4%) or Unknown/ Undeclared caste (7.2%).[7]

   Notably, a prominent woman politician from marginalised caste received significantly more caste-based slurs than others. This indicates that caste identity is more often than not, a key element of problematic or abusive content for women belonging to marginalised castes.

7. **Women politicians from political parties other than the Bharatiya Janata Party experienced more abuse**

   While women politicians across all political parties experienced sexist abuse, their overall experience was divided along party lines.

   As compared to women politicians associated with the Bharatiya Janata Party (BJP), which is also the ruling party in India currently, women politicians from 'other parties' [8] experienced 56.7% more problematic or abusive content. Women politicians associated with the Indian National Congress (INC) also received 45.3% more problematic or more abusive content than BJP.

Online abuse against women on this scale should not and does not have to exist on social media platforms. Companies like Twitter have a responsibility to respect human rights, which means ensuring that women and marginalised groups using the platform are able to express themselves freely and without discrimination.

In the recent past, Twitter has admitted that it has created an unsafe space for women by perpetuating harassment and abuse. While Twitter has guidelines in place to identify abuse and hate, and has also improved its policies and reporting process over the years, the findings suggest that these policies are not sufficient to address the toxicity that women face online.

Amnesty International through its Toxic Twitter study (2018)[9] and Troll Patrol study (2018)[10] has repeatedly asked Twitter to make available meaningful and comprehensive data regarding the scale and nature of abuse on their platform, as well as how they are addressing it.

Based on the findings of this study, we recommend further steps for Twitter to ensure that its policies are transparent, uniform, and are based on human rights standards and gender-sensitive due diligence. Considering India's linguistic diversity, Twitter should ensure coverage not only of India's main languages, but also regional languages, with due focus on mixed language tweets where native scripts are used alongside Latin scripts. Further, it should ensure that discrimination and abuse on the basis of gender, caste, religion, ethnicity, gender identity, sexual orientation and other identifying factors does not prevent users from exercising their right to freedom of expression equally on the platform. Importantly, it must constantly and transparently evaluate and measure whether it is effectively tackling online violence against women.

Online abuse has the power to belittle, demean, intimidate and eventually silence women. Twitter must reaffirm its commitment to providing a 'safe space' to women and marginalised communities. Until then, the silencing effect of abuse on the platform will continue to stand in the way of women's right to expression and equality.

"People should know what women in politics endure, what they have to put up with and how unequal it becomes for them. It is such a tough battlefield, so to speak. Really I do believe that Twitter is my workplace."

"But if my workplace were to be a battlefield, all the time, would I be able to contribute, to the cause that I represent, easily and with fairness, if I am constantly being attacked for being a woman."

– *Shazia Ilmi, Member, Bharatiya Janata Party*

[6] Amnesty International, Troll Patrol Findings, https://decoders.amnesty.org/projects/troll-patrol/findings

[7] Marginalised caste, in this study, included politicians belonging to Scheduled Caste, Scheduled Tribe and Other Backward castes. The Scheduled Caste (SCs) and Scheduled Tribes (STs) are officially designated groups of historically disadvantaged people in India. Articles 341 and 342 of the Constitution of India define who would be Scheduled Castes and Scheduled Tribes with respect to any State or Union Territory. For the purpose of study, the caste identity of the women was taken from *Lok Dhabha*, an initiative of Ashoka University. 'Unknown/ Undeclared caste' refers to those who either did not publicly declare their caste and/or Amnesty International India was unable to identify them with a particular caste group.

[8] Other parties included Aam Aadmi Party, Apna Dal, All India Anna Dravida Munnetra Kazhagam, All India Trinamool Congress, Bahujan Samaj Party, Community Party of India (Marxist), Communist Party of India (Marxist-Leninist), Dravida Munnetra Kazhagam, Jammu & Kashmir National Congress, Jammu & Kashmir Peoples Democratic Party, Jharkhand Mukti Morcha, Rashtriya Janata Dal, Shiromani Akali Dal, Shiv Sena, Samajwadi Party, Telangana Rashtra Samiti, Yuvajana Sramika Rythu Congress Party. Our analysis could not be further disaggregated by 'other' parties as we did not have statistically relevant samples for each party (having a sample per party between 1 and 4 members). Further, we observed that, except for some notable cases, many women from these parties did not have an active Twitter account.

[9] Amnesty International, Toxic Twitter, INDEX NO. ACT 30/8070//2018

[10] Amnesty International, Troll Patrol Findings, https://decoders.amnesty.org/projects/troll-patrol/findings

Methodology

Troll Patrol India project is a joint effort by human rights researchers, technical experts and thousands of online volunteers to build a large crowdsourced dataset of online abuse against women politicians in India.

This chapter explores the methodology adopted to measure the scale and nature of online abuse faced by women politicians in India. It discusses 'who did we study', 'when did we study' and 'how did we study'.

The women politicians represented a variety of political views spanning the ideological spectrum. Using data science we were able to provide a quantitative analysis of the scale of online abuse against women politicians in India.

Troll Patrol India is a crowdsourcing effort to demonstrate the scale and nature of online abuse against Indian women politicians on Twitter in the context of the 2019 General Indian Elections of India. Following the methodology developed together with Element AI for Amnesty International's 2018 Troll Patrol study, we built a database of over 114,716 tweets mentioning 95 women politicians from India and asked digital volunteers, known as 'Decoders' to identify problematic and abusive content in those tweets.

An impressive number of 1,912 Decoders from 82 countries analysed the tweets to create a labelled dataset of problematic or abusive content. The Decoders were shown a tweet with username obscured, mentioning one of the women in our study, then were asked simple questions about whether the tweets were problematic or abusive, and if so, whether they revealed misogynistic, casteist or racist abuse, or other types of violent threats. Each tweet was analysed by multiple Decoders.

The Decoders were shown a video tutorial and definitions and examples of problematic and abusive content, and they were encouraged to engage on an online forum where they could discuss the tweets with each other and with Amnesty International's and Amnesty International India's researchers.

In total the Decoders labelled 142,474 tweets, totalling 474,383 individual answers. This large number allowed us to have both a dataset of 114,716 tweets sampled uniformly at random, and a smaller dataset of 27,758 tweets sampled from a more experimental procedure. The experimental dataset was made possible by the huge amount of interest in the campaign. It was a learning opportunity for us to further our experience of using advanced analysis tools on real world data, and preparing the terrain for future campaigns. In the interest of simplicity and methodological clarity, however, all the statistics in the rest of this report were computed on the basis of the tweets sampled uniformly at random.

Using the subset of 114,716 tweets sampled uniformly at random and annotated by the Decoders, we extrapolated the abuse analysis to the full 7 million tweets that mentioned the Indian women politicians selected for our study. The results published in this study are based on this.

The tweets were deployed through Troll Patrol India page[11] - hosted on Amnesty Decoders,[12] the micro-tasking platform based on Hive (Labs, 2014) and Discourse (Discourse). This platform by Amnesty International engages Decoders (mostly existing members and supporters) in human rights research.

The Troll Patrol India microsite was launched for decoding on 15 May 2019 and was open till 8 August 2019. Great effort was put into designing the user-friendly and interactive interface, and could be accessed by Decoders through computers and mobile phones.[13]

STUDIED POPULATION: WOMEN POLITICIANS ON TWITTER

IDENTIFYING INDIAN WOMEN POLITICIANS

The sample of women politicians was identified prior to the submission of nomination papers by the women contesting in the 2019 General Elections. The criteria for sample selection were as follows-

- Members of Parliament in the two most recently elected Lower House of Parliament (15th and 16th Lok Sabha)
- Members of Parliament in the two most recently elected Upper House of Parliament (Rajya Sabha)
- Members of the Legislative Assembly of the States and Union Territories as of February 2019
- Party office bearers and spokeswomen for all the national and regional political parties
- Members from reserved constituencies (seats reserved for specific groups in the Indian Parliament at national and state level)
- Current and former chief ministers (elected heads of government) for all states and union territories

The Twitter accounts of these women were identified. If the account was not "verified" (a Twitter badge that recognises the account to be authentic)[14] then the team investigated to ensure that the account was genuine, by using party websites and other such sources. Our research resulted in 101 Twitter handles. Of these some had very little Twitter activity – 95 politicians had at least one tweet mention in the final dataset and 82 had 10 or more tweet mentions in the sample set given to Decoders.

[11] Troll Patrol India, Amnesty International India, https://decoders.amnesty.org/projects/troll-patrol-india

[12] Troll Patrol, Amnesty International, https://decoders.amnesty.org/projects/troll-patrol

[13] *See,* Annexure 3, Amnesty Decoders Tool and Screenshots

[14] Note that Twitter verification has been mostly on hold since 2018. *See*, About Verified Accounts, Twitter, https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts

PROFILE OF THE POLITICIANS

We researched characteristics of these politicians such as party affiliation, political inclination, their official role prior to the elections, current post, whether they contested the election, whether they won the election, their religion, caste, age, sexual orientation, gender identity, marital status, their twitter followers and their twitter mentions. For this classification, we used official sources such as ethnic diversity studies and information in the public domain such as public profiles, articles written by or about these women, official websites including party websites, Parliament websites, nomination papers, *Lok Dhaba* and Wikipedia pages. It is important to note that this is an approximate classification for the purpose of this analysis, utilising research available in public domain and is not necessarily a reflection of how each politician identifies herself.

SELECTING A TIMELINE

We aimed to study tweets directed at women politicians during the three-month period of March-May 2019 - that is, in the lead-up to, during, and shortly after the 2019 General Elections of India. Therefore, tweets mentioning women politicians over these 3 months, were sampled. The General Elections ran from 11 April 2019 to 19 May 2019, the results of which were announced on 23 May 2019.

OBTAINING AND PROCESSING TWITTER DATA

OBTAINING SAMPLE TWEETS

The women in this study were mentioned in over 7 million tweets between March and May 2019. Crimson Hexagon (now part of Brandwatch) was used to obtain a subsample of these tweets. A random sample of 10,000 tweets were extracted per day as per the Twitter API (Application Program Interface) terms. The sample included tweets mentioning at least one of the women politicians' twitter handles and excluded retweets and tweets that were deleted before the date of extraction.

Although the API allows 10,000 tweets per day, the actual number was lower due to exclusion of deleted or private tweets and retweets. We also removed all completely content-free tweets (those that were just a list of @mentions and no other content) – such tweets amounted to 1.7% of our initial sample.

Out of the data obtained from Crimson Hexagon, we sampled 142,474 tweets for the Troll Patrol India project. The sample that was labelled by Decoders included 114,716 tweets sampled uniformly at random, and a smaller dataset of 27,758 tweets sampled from a more experimental procedure.[15] All the statistics in the rest of this study were computed on the basis of the 114,716 tweets sampled uniformly at random.

HANDLING TWEET LANGUAGES

Considering India's linguistic diversity, this study also looked at tweets in languages other than English. Based on an initial sample, Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Tamil and Telugu were found be the most commonly used languages in the tweets mentioning the women politicians. Keeping this in mind, at the time of registration, decoders were given the option of choosing one or more languages from these nine languages for decoding.

We could not obtain the language classification from Twitter itself through the third-party service used to access the data. We instead, detected languages using an automated language detection tool (Google Inc.). Some of the challenges in this approach included, very short content for some tweets, no access to any user data to estimate language, the ability of users to use both Latin and native scripts on Twitter, the usage of 'Hinglish' – a mix of Hindi and English – or other mixtures of languages, etc. Despite its shortcomings, the classification was useful in engaging Decoders in their own language and in providing a rough analysis of abuse in different languages. Importantly, using the available tools, the Decoders could provide feedback on language mischaracterisation.[16]

DECODING - ANALYSIS OF ABUSE BY VOLUNTEERS

The tweets obtained through random sampling were analysed through a process called 'Decoding'. Amnesty Decoders is an innovative platform for volunteers around the world to help Amnesty researchers in analysing large volumes of pictures, documents and social media messages through their computers and phones. These Decoders were trained in identifying and categorising problematic and abusive content. After a video tutorial, Decoders were asked to select their language(s) of preference. After being shown a tweet from an obscured username, which mentioned one of the women in our study, they were asked multiple-choice questions:

1. "Does the tweet contain problematic or abusive content?" (No, Problematic, Abusive).
2. Unless their answer was No, the follow-up question

was "What type of problematic or abusive content does it contain?" The Decoders could select one or more responses from a list, namely - sexism or misogyny, ethnic or religious slur, racism, casteism, homophobia or transphobia, physical threat or sexual threat or a miscellaneous 'other' category for those tweets that could not be easily categorised.

3. The Decoders had to further categorise the medium of abuse - whether in the form of text, image or a video.

The Decoders could, at any point in time, access the definitions and examples of problematic and abusive content.[17]

EXPERT ANALYSES

Apart from engaging the Amnesty Decoders, we also had 796 tweets assessed in-house by Amnesty International India's staff familiar with the parameters of 'problematic' and 'abusive' content. Each tweet was analysed by three experts, thus enabling the validation of response. For each of the nine languages, three different experts looked at all of the tweets in that particular language, in order to allow for an agreement analysis between the experts. People fluent in a particular language decoded tweets of that language. For example, 261 tweets in English were decoded by three experts fluent in English, while 38 tweets in Marathi were decoded by another three experts fluent in Marathi.[18]

GENERATED ESTIMATES OF PROBLEMATIC AND ABUSIVE CONTENT

AGREEMENT ANALYSIS

The Amnesty team investigated whether Decoders generally agreed on the classification of a tweet as problematic or abusive. Due to the subjective nature of the analysis, we did not expect perfect alignment between individuals. We also compared the agreement amongst the crowd and experts. We expected "expert" Amnesty International India staff familiar with the issues to have higher agreement than the "crowd" of Decoders.

Agreement amongst raters of tweets was quantified using two measures: Fleiss' kappa and intra-class correlation coefficient.[19]

There was fair agreement of experts on whether tweets contained problematic, abusive or neutral content and good agreement when the ordinal nature of the classes was considered (that is, 'Problematic' and "Abusive' answers were closer together than 'Abusive' and 'No').

When grouping together the 'problematic' and 'abusive' classes, expert agreement was moderate. Expert agreement with respect to the type of 'problematic' or 'abusive' content was also fair. As was expected, agreement of Decoders was lower than that of experts.

Agreement amongst Decoders was fair when the ordinal nature of the classification was taken into consideration. Decoder agreement was also fair when grouping together the 'problematic' and 'abusive' classes and also when evaluating the type of content for tweets classified as such. The analysis showed that our crowd results were better than random chance.[20]

AGGREGATED PROPORTIONS ACROSS TWEETS

For the descriptive statistics given in the findings of this study, we used the same methodology developed by Amnesty International and Element AI for the 2018 Troll Patrol Study[21], aggregating Decoders' annotations ("votes") across all tweets in the set, as opposed to aggregating the majority vote.[22]

Our method of treating each vote as a measure meant that the following two scenarios on 10 tweets would lead to the same aggregated estimate of 30% abuse (simplified as "3 abusive tweets"):

[15] *See,* Annexure 1, Enrichment

[16] For more information about handling languages, *see,* Annexure 2

[17] For definitions and examples of problematic and abusive tweets, see, chapter on Methodology, page 14-15

[18] For results of the expert decoding agreement analysis, *see*, Annexure 5.

[19] (Fleiss, 1971) and intra-class correlation coefficient (Shrout, 1979). https://psycnet.apa.org/buy/1972-05083-001

[20] *See,* full agreement analysis and methodology, including definitions of fair and moderate agreement in Annexure 5.

[21] Amnesty International, Troll Patrol Findings, https://decoders.amnesty.org/projects/troll-patrol/findings

[22] Laure Delisle et al, Troll Patrol Methodology Note, https://decoders.azureedge.net/data-viz/images/Troll%20Patrol%20-%20Methodology.pdf

- 10 tweets, each labelled as abusive by 30% of its raters, and as neutral by the remaining 70% of its raters.
- 10 tweets, 3 of which were labelled as abusive by 100% of their raters, and the other 7 are labelled as neutral by 100% of their raters.

This was justified by the agreement analysis discussed in the previous section and Annexure 5. Unlike a visual classification task with a clear and objective ground truth, raters were often divided, especially when the abuse in a tweet was subtle or contextual. The aggregation of proportions accounts for this variability of opinions and reflected the most accurate picture, without resorting to the extreme solution of a majority vote, which dismissed the minority's perception of abuse, or considered a tweet abusive if at least one of its raters voted as such, which overestimated the overall perception of abuse.

WEIGHTING

The answers of the Decoders were re-weighted to account for sample mismatch with the pool of tweets (the true distribution in the "world set"). Weighting factors used for both responses and tweets included: 1) Response weighting accounting for tweets that were analysed less or more than the expected number of times (e.g. if one tweet had two out of six annotations as Abusive, and one had one out of three annotations as Abusive, those contributed in the same way); 2) Weighting by tweet collection 'batch' used a uniform sampling per day (e.g. if one date had 20,000 tweets mentioning the women and a second date had 10,000, the first date should have had twice as many tweets in the sample) and 3) Weighting by tweet language distribution accounting for a difference of completion between different languages for the last batch of data.[23]

ACCOUNTED FOR SAMPLING UNCERTAINTY: BOOTSTRAPPING

Every statistical analysis based on a sample must evaluate the robustness of its findings against the randomness induced by its sampling mechanism. In this study, one major but controllable source of uncertainty stemmed from the limitation in data collection: Maximum 10,000 tweets per day, sampled by the provider uniformly at random, among all the tweets of that day mentioning the politicians identified. This is but a small portion of the much larger volume of such tweets from that day.

In an ideal world with unlimited resources, we would have measured the robustness of our findings by carrying out the research project many times over, including resource intensive steps of data collection and crowdsourced labelling, and comparing the results. However, given the actual unfeasibility of this approach, we instead obtained measures of accuracy of the results through the statistical method of bootstrapping.

Bootstrapping involved taking 100 random resamples with replacement from the weighted set of 114,716 labelled tweets. Each random resample taken was of the same size as the set of labelled tweets and resampling with replacement implied that a specific tweet may have ended up multiple times in a single resample whereas another tweet may not have ended up in the same resample at all.

The frequencies of tweets with problematic or abusive content as well as the frequencies of types of problematic or abusive content presented in this report were calculated as the medians across the 100 resamples. The uncertainty associated with these frequencies and conversely thus the robustness of our findings were provided through 95% confidence intervals, which were calculated as the range between the 2.5% and 97.5% quantiles across the 100 resamples. The 95% confidence intervals implied, roughly speaking, a 95% level of confidence that the intervals contained the true frequencies over all of the more than 7 million relevant tweets.

In short, bootstrapping entailed getting multiple estimates of the frequencies, each on a random resample with replacement from the actual set of labelled tweets, and aggregating over the resamples to gauge the accuracy of our findings. The rationale for employing the bootstrapping was that repeated sampling from the empirical distribution of the labelled tweets was the closest approximation to the aforementioned ideal but unfeasible repeated sampling from the real-world distribution of the over 7 million tweets by conducting the research project.

LIMITATIONS OF THE STUDY

Crude method used for Language detection: Due to limited resources, we used a very basic method for language detection, namely Google's language auto detection. This method resulted in the language of some tweets being wrongly identified, which could have reduced users' experience as they may have received tweets in a language they had not selected.

English as the preferred language for Decoding: A high proportion of the Decoders selected English as their primary language, possibly limiting the diversity of the study.
This could be a result of the recruitment efforts that were primarily in English and the limitations in our outreach as

an organisation. This also resulted in English tweets being exhausted quickly, and lesser engagement on the tweets in other languages.

Exclusion of first-time politicians: Women politicians who contested elections for the first time were excluded in this study, since at the time of creating the sample of politicians for study, the information on their candidature or nomination was either not final or available online. Some of these women experienced online abuse during their campaign and after elections, and the study would have benefitted from their inclusion.

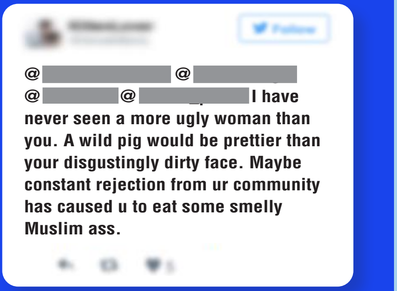
Tweet sampling and de-identifying: Decoders analysed random tweets with Twitter handles blurred. We chose to de-identify the tweets to protect individuals sending and receiving problematic and abusive content. This meant that Decoders could not tell if a woman politician was being targeted herself or merely included as one of a list of names. The tweets were also shown without the corresponding thread or conversation due to limited access to Twitter data. This meant that Decoders could not use the rest of a conversation to judge a tweet.

Perception of Decoders: The study has relied on the "perception of the Decoder", which is influenced by their experiences and prejudices. What may be considered as problematic or abusive to one may not be considered the same by another. The project addressed this by showing the same tweet to many Decoders, which facilitated "validation" of responses, but it is still possible that “atypical” opinions have influenced our findings.

Timeline: It is important to note that the findings of this study are restricted to the time period during which the tweets were sampled i.e. between 1 March 2019 to 31 May 2019. While the elections were held from 11 April 2019 to 23 May 2019, the period between March and May witnessed extensive election campaigning. A longer period of tweet samples may have enabled the study to capture the abuse experienced by women politicians over a longer period of time. It would have also shed light on the patterns of abuse during normal and peak times, depending on political events and campaigning around the year.

[23] *See,* Annexure 4, Weighting

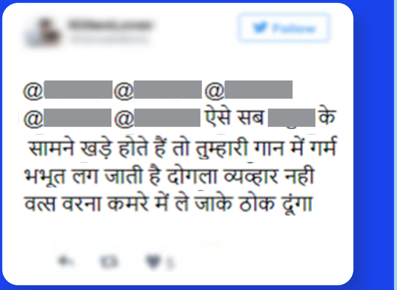
DECODING TWEETS - DEFINITIONS



[redacted]

@[redacted] @[redacted] @[redacted] @[redacted] I have never seen a more ugly woman than you. A wild pig would be prettier than your disgustingly dirty face. Maybe constant rejection from ur community has caused u to eat some smelly Muslim ass.

'Problematic'- Problematic tweets contain hurtful or hostile content, especially if it were repeated to an individual on multiple or cumulative occasions, but not as intense as an abusive tweet. It can reinforce negative or harmful stereotypes against a group of individuals (e.g. negative stereotypes about a race or people who follow a certain religion). Such tweets may have the effect of silencing an individual or groups of individuals.

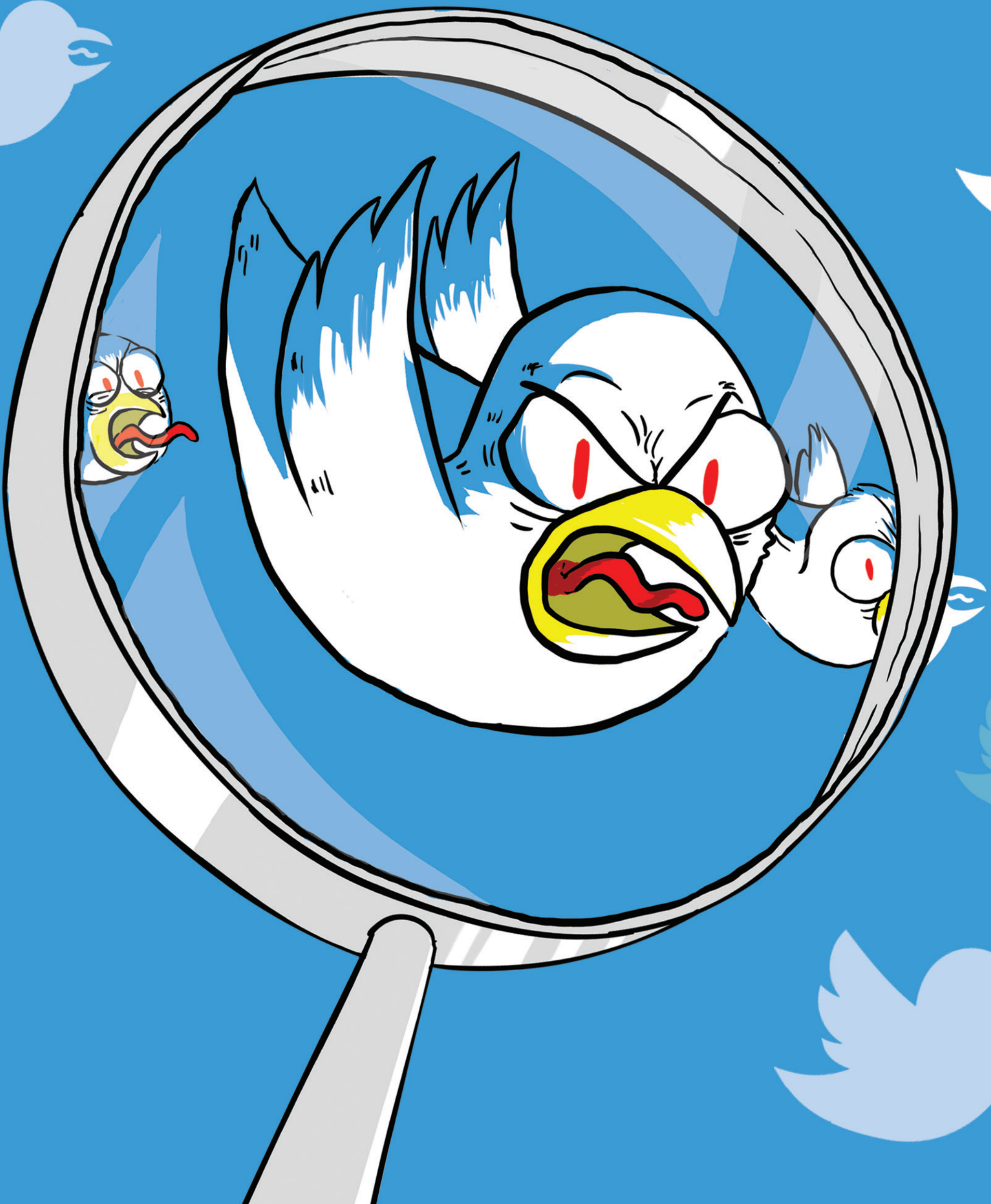


[redacted]

@[redacted] @[redacted] @[redacted] @[redacted] @[redacted] ऐसे सब [redacted] के सामने खड़े होते हैं तो तुम्हारी गान में गर्म भभूत लग जाती है दोगला व्यव्हार नही वत्स वरना कमरे में ले जाके ठोक दूंगा

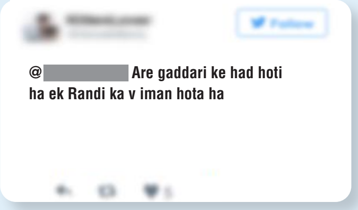
'Abusive' – Abusive content violates Twitter's own rules and includes tweets that promote violence against or threaten people based on their race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.

Examples include: physical or sexual threats, wishes for the physical harm or death, reference to violent events, behaviour that incites fear or repeated slurs, epithets, racist and sexist tropes, or other content that degrades someone.



WARNING: THE EXAMPLES BELOW CONTAIN EXPLICIT AND THREATENING MESSAGES THAT SHOW VIOLENCE AGAINST WOMEN.

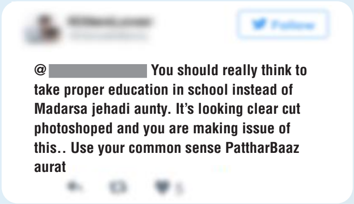
The @twitterhandles and @mentions have been blurred since the purpose here is to highlight the nature of abuse and not the persons who are abused, nor the persons who abuse.



@[redacted] Are gaddari ke had hoti ha ek Randi ka v iman hota ha

SEXISM OR MISOGYNY

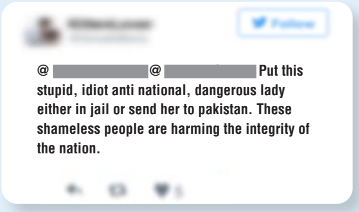
Insulting or abusive content directed at women based on their gender, including content intended to shame, intimidate or degrade women. It can include profanity, threats, slurs and insulting epithets.



@[redacted] You should really think to take proper education in school instead of Madarsa jehadi aunty. It's looking clear cut photoshoped and you are making issue of this.. Use your common sense PattharBaaz aurat

ETHNIC OR RELIGIOUS SLUR

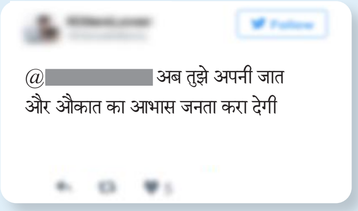
Discriminatory, offensive or insulting content directed at a woman based on her religious beliefs and or ethnicity, including content that aims to attack, harm, belittle, humiliate or undermine her and her community.



@[redacted] @[redacted] Put this stupid, idiot anti national, dangerous lady either in jail or send her to pakistan. These shameless people are harming the integrity of the nation.

RACISM

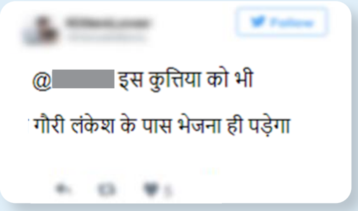
Discriminatory, offensive or insulting content directed at a woman based on her race, including content that aims to attack, harm, belittle, humiliate or undermine her.



@[redacted] अब तुझे अपनी जात
और औकात का आभास जनता करा देगी

CASTE SLUR

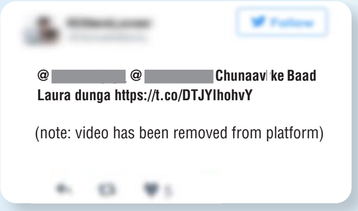
Discriminatory, offensive or insulting content directed at a woman based on her caste, including content that aims to attack, harm, belittle, humiliate or undermine her or her community.



@[redacted] इस कुत्तिया को भी
गौरी लंकेश के पास भेजना ही पड़ेगा

PHYSICAL THREATS

Direct or indirect threats of physical violence or wishes for serious physical harm, death, or disease.

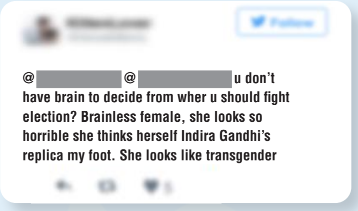


@[redacted] @[redacted] Chunaav ke Baad Laura dunga https://t.co/DTJYIhohvY

(note: video has been removed from platform)

SEXUAL THREATS

Direct or indirect threats of sexual violence or wishes for rape or other forms of sexual assault.



@[redacted] @[redacted] u don't have brain to decide from wher u should fight election? Brainless female, she looks so horrible she thinks herself Indira Gandhi's replica my foot. She looks like transgender

HOMOPHOBIA OR TRANSPHOBIA

Discriminatory, offensive or insulting content directed at a woman based on her sexual orientation, gender identity or gender expression. This includes negative comments towards bisexual, homosexual and transgender people.

OTHER

There will be some tweets that fall under the 'other category' that are problematic and/or abusive. For example, statements that target a user's disability, be it physical or mental, or content that attacks a woman's nationality, health status, legal status, employment, etc.

Decoders: The Spirit of Volunteerism Enabled the Crowdsourced Research

The research on this scale by Amnesty International India and Amnesty International could only have been possible because of the volunteering spirit demonstrated by civil society, which contributed both time and efforts to the process generously.

This chapter explores the profile of the Decoders.

DECODERS BACKGROUND: COUNTRIES AND CITIES

A total of 1,912 Decoders were involved in Troll Patrol India, representing 82 countries (based on registration). Most Decoders (57.3%) came from India (based on user-declared country during registration). For most users (1,764 of 1,912) this study was their first Decoders project. Most returning users were from outside of India, presumably (and anecdotally from forum comments) joining the project out of general interest in volunteering.

Decoders by country of origin

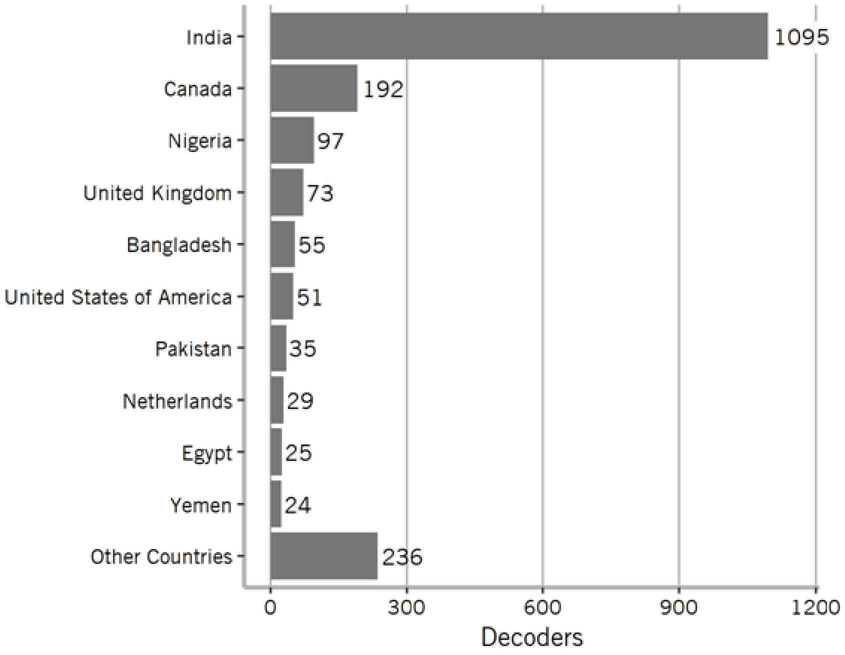


Fig: 1

Within India, most Decoders engaged from Delhi. While the second highest number of Decoders were from Uttar Pradesh, within the state most Decoders came from Noida which is a part of the Delhi National Capital Region (Delhi NCR), making Delhi NCR the highest source of volunteering in this project.

State of origin for Indian Decoders

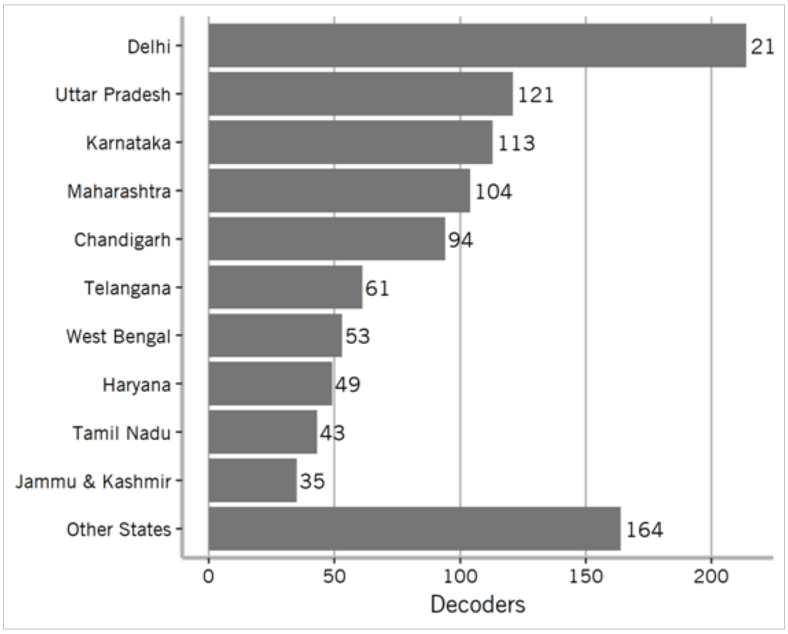


Fig: 2

DECODING SUBMISSIONS

Decoding took place between 13 May 2019 and 8 August 2019. Promotional events at colleges in Delhi and Punjab and the 'Decodathon' in Bengaluru resulted in spikes of decoding rate.

DECODER BEHAVIOUR

The median assignments completed by a Decoder was 24 and the median time spent in a decoding session was 15.5 minutes. This is good in terms of engagement with a voluntary online task.

Of the 1,912 registered Decoders, 88% (1,692) decoded, while the rest browsed and/or quit for various possible reasons.

USER LANGUAGE

Most common language option for Decoders

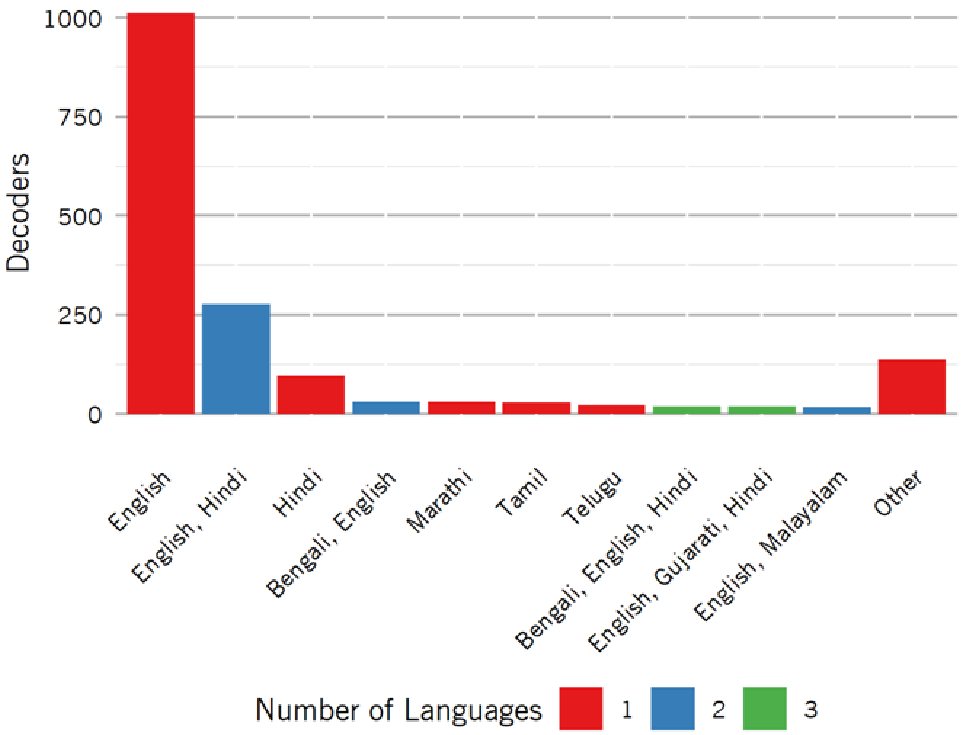


Fig: 3

Common choice of language for Decoders: By far, English was the most-opted language by users (1,012 users). In fact, most users decoded tweets in English at some point (1,464 or 86.7%), regardless of what other language was selected.

This could be attributed to English being the main language of the Decoders website, including the help taskbar and the registration process. Although the training video for Troll Patrol India was also translated into Hindi, which is the most commonly tweeted language in India - and examples of Hindi tweets were also provided - some knowledge of English, by default became a baseline requirement. Most of our earlier projects have been in English, therefore anyone volunteering due to previous interest was more likely to be an English-speaker. Lack of time and resources for the translation of the Decoders website into other languages also contributed to the use of English.

We also acknowledge that these choices may have led to reduced diversity in the outlook of the Decoders assessing tweets. Further, their opinions may not be representative of the wider Indian population about what would count as problematic or abusive content, and views of some cultural groups - especially those less likely to be fluent in written English.

ENGAGEMENT OF DECODERS

Decoders were directed to the website through a combination of online and offline engagement.

ONLINE ENGAGEMENT

Decoders were engaged digitally through advertisements on social media and targeted emails that were sent out to existing and new supporters, encouraging them to sign up for this project. Twenty volunteers were engaged, using online platforms such as *Internshala, Lawoctopus, Latestlaws* and *Noticeboard*. While these volunteers decoded tweets themselves, they also engaged other volunteers.

Two volunteers were engaged to monitor and respond to comments on the digital community forum and to engage with Decoders online between May 2019 and July 2019. These volunteers also categorised thousands of flagged tweets based on language.

OFFLINE ENGAGEMENT

Offline engagement was based on a model where we engaged volunteers who in turn mobilised other volunteers. Events were conducted in Delhi and Punjab to engage Decoders, of whom most were students.[24] They were briefed on the project and were given an orientation on decoding. They were asked to read tweets mentioning women politicians on Twitter and then asked to analyse the tweets, as part of a mock decoding exercise.

A melange of offline and online engagement was used to organise a 'Decodathon' in Bengaluru on 6 July 2019. The event was promoted through a Facebook event page and emails to our Bengaluru-based volunteers. During the event, speakers engaged the audience and the participants debated on the reasons behind trolling in India.[25]

DECODER SURVEY

After analysing 10 tweets, each decoder was given a voluntary survey. The questions related to basic biographic details about the Decoders and their past experience of facing online abuse.

55.9% or 946 of the Decoders answered the survey. More than half of the Decoders who took the survey, identified as female.

Both female and male Decoders reported past experience of online abuse. We recognised that this was a limited survey and did not indicate the nature, degree or sources of online abuse, but this indicated the situation of underreporting of abuse by men. While 35.4% (677) of all Decoders reported experiencing abuse, 14.5% (287) were unsure.

AGE OF THE DECODERS

The majority of Decoders in Troll Patrol India who completed the survey (76%) were under 34 years of age. This reflected Amnesty International India's success in targeted engagement of university students, which included holding interactive events at colleges.

[24] A total of 565 people attended the events. The breakup is as follows: Khoj Studio (35 attendees), BR Ambedkar College, Delhi University (80 attendees), Punjab University, Chandigarh (55 attendees), Maitreyi College, Delhi University (72 attendees), Hindu College, Delhi University (100 attendees), Kirori Mal College, Delhi University (100 attendees), Kamala Nehru College, Delhi University (108 attendees) and Indian Institute of Technology, Delhi (15 attendees)

[25] You can read more about the event at https://amnesty.org.in/decodathon-indiatrollpatrol-irl-in-real-life/

Findings

“The online abuse that followed in 2014 for a young girl entering politics was traumatising. The trolling was sexist, misogynistic and targeted me for being Muslim. I was told that *‘I have no right to speak as a Muslim woman’*. Rape threats were routine, as were character assassinations, insinuations about my sexual relationships with older men”.

"Now in 2019, I have considerably reduced my activity on Twitter. I ask myself how trollable is that and whether I really need to put up my opinion.”

– Hasiba Amin, National Convener, Social Media, Indian National Congress

1. 1 IN EVERY 7 TWEETS THAT MENTIONED WOMEN POLITICIANS IN INDIA WAS 'PROBLEMATIC' OR 'ABUSIVE'

> "If you were to see the trolling handles you will see a common political ideology spanning across the handles. They don't refrain from hitting out on women from their own party if they feel they have been out of line."
>
> *– Atishi, Member, Aam Aadmi Party, Political Affairs Committee & National Executive Advisor, Deputy Chief Minister, Government of Delhi*

Our study found that **13.8%** of tweets mentioning the women in the study were problematic or abusive. This amounted to nearly **1 million problematic or abusive mentions of these 95 women** across three months around the 2019 General Elections of India. On average, this amounted to over 10,000 problematic and abusive tweets every day across all women in the sample, or 113 per woman per day.

Frequency of problematic and abusive content

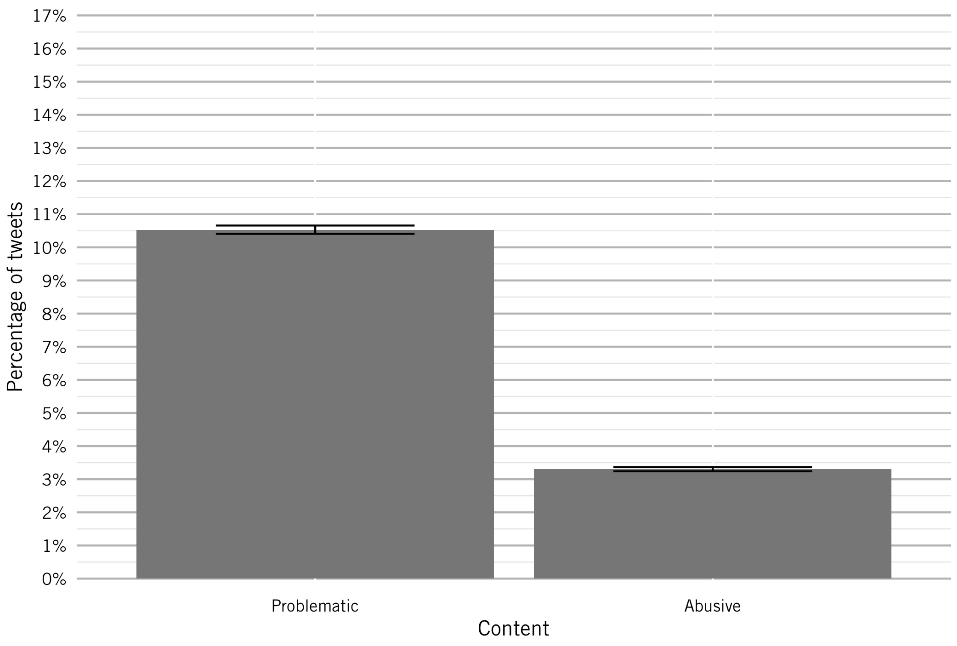


Fig: 4

Problematic	Abusive
10.5%	3.3%

2. INDIAN WOMEN POLITICIANS EXPERIENCED SUBSTANTIALLY HIGHER ABUSE THAN THEIR UK AND USA COUNTERPARTS

A similar study conducted by Amnesty International in the UK and the USA in 2018 to measure online abuse faced by 323 women politicians found that 7.1% tweets mentioning politicians were problematic or abusive. Whereas the Troll Patrol India study, using a similar methodology, but focusing on a shorter period during elections found that Indian women politicians experienced 13.8% problematic or abusive tweets, which was substantially higher.

Abuse against women politicians in UK, USA and India

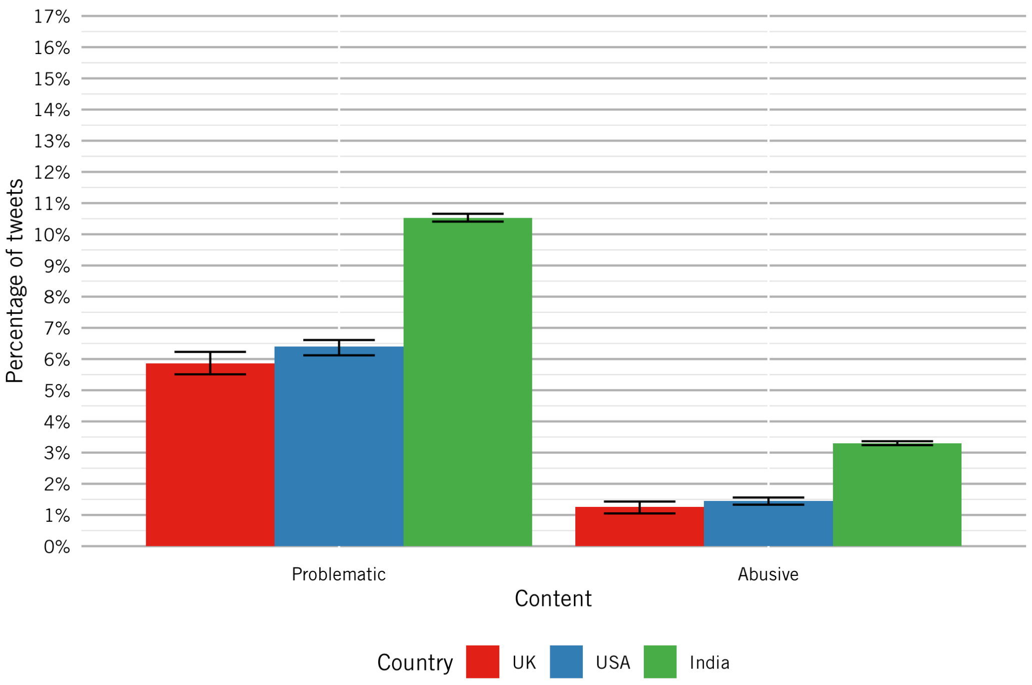


Fig: 5

Notes:

- *Abuse against women politicians in the UK and the USA was measured in the Troll Patrol study for a period of one year (2017). Abuse in India was measured through the Troll Patrol India study using a similar methodology but focusing on a shorter period during elections (March-May 2019). This may account for some of the difference in frequencies.*

Country	Problematic	Abusive
UK	5.9%	1.3%
USA	6.4%	1.5%
India	10.5%	3.3%

3. WOMEN POLITICIANS WHO ARE PROMINENT ON TWITTER WERE TARGETED MORE

We wanted to analyse if politicians who are more prominent on Twitter (i.e. who receive a higher amount of mentions) received more abuse.

Proportion of abuse by number of mentions

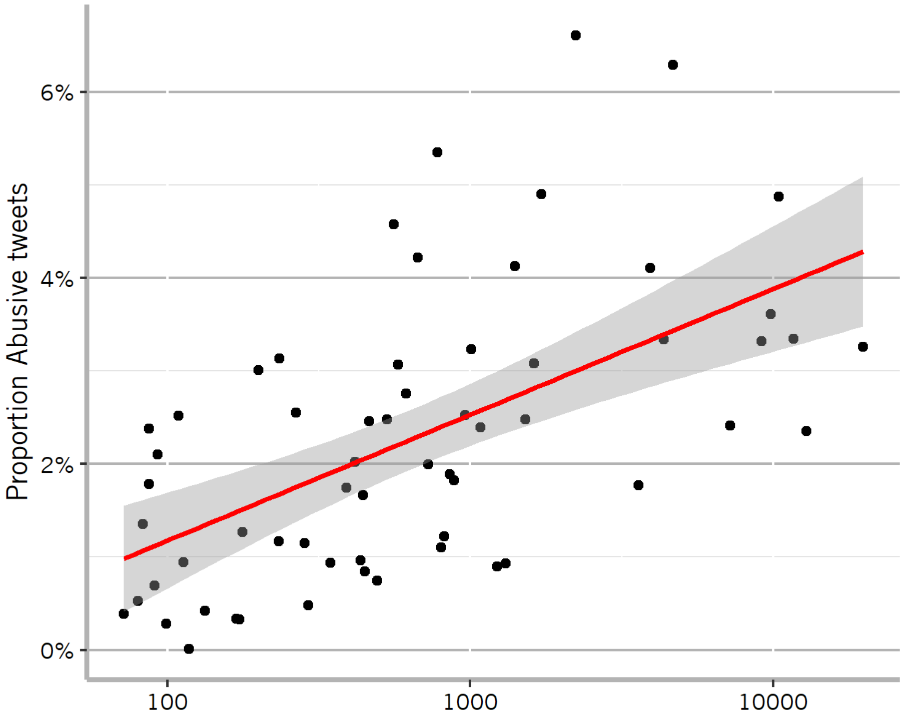


Fig: 6

Notes:

- *This chart includes politicians with over 50 mentions in the sample*

There is a clear positive association between the number of mentions and the proportion of abusive content received. This finding confirms that being more visible on social media as a woman politician can lead to being targeted more for abuse.

We also confirmed this by grouping the politicians by mentions. To do this, we split the group into the top 10 most-mentioned politicians and the rest, and observed that the top 10 politicians had a mean of 14.8% problematic or abusive content vs 10.8% for the others. This meant that the top 10 received 74.1% of all mentions, but 79.9% of problematic or abusive mentions.

Comparison of top 10 most mentioned politicians vs others

	Problematic or Abusive tweets	Proportion of tweet mentions	Proportion of abuse
Top 10 most mentioned politicians	14.8%	74.1%	79.9%
All other politicians	10.8%	25.9%	20.1%

4. 1 IN EVERY 5 PROBLEMATIC OR ABUSIVE TWEETS WAS SEXIST OR MISOGYNISTIC

On the Troll Patrol India website, when Decoders identified a tweet as problematic or abusive, they were also asked to identify the type of 'problematic' or 'abusive' content, from the choice of categories as shown on website. These categories were sexism or misogyny, ethnic or religious slur, racism, casteism, homophobia or transphobia, physical threats, sexual threats and/or other. They could choose more than one category. Each category had definitions and examples available on the Help section of the website. Brief description of the categories was also made available which could be accessed by hovering the cursor over the information icon next to each category.

The categories of types of 'problematic' or 'abusive' content has been detailed in chapter on Methodology. Examples of tweets have also been shared.

For instance, the information icon for ‘caste slur’ defined it as “discriminatory, offensive or insulting content directed at a woman based on her caste, that aims to attack, harm, belittle, humiliate or undermine her and her community.”

The most common selection for type of abuse was the 'other' category (74.1%), suggesting most problematic or abusive content did not fit neatly into one of the suggested categories in a way users were comfortable with.

Frequency of type of problematic or abusive content

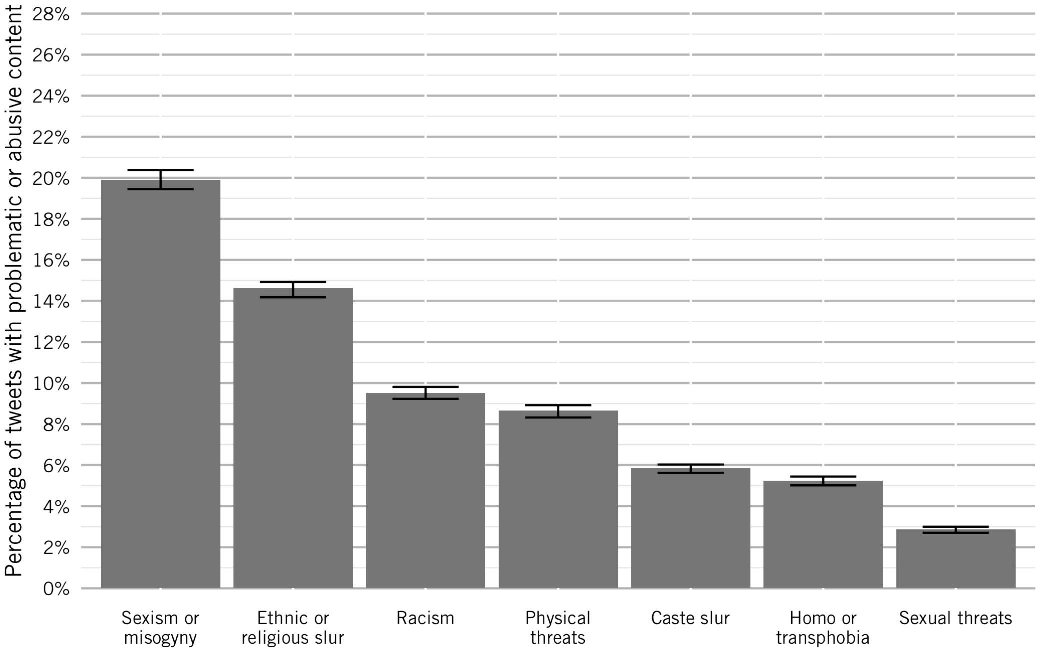


Fig: 7

Sexism or misogyny	Ethnic or religious slurs	Racism	Physical threats	Caste slurs	Homophobia or transphobia	Sexual threats
19.9%	14.6%	9.5%	8.7%	5.9%	5.2%	2.9%

Nearly one in every five problematic or abusive tweets (19.9%) displayed sexism or misogyny, according to the selection of the decoders.

> “There are abusive tweets and messages about the colour of my skin, how ugly I am, too ugly to be raped, and lots of tweets about my private parts, and all kinds of ugly abuse combining misogyny and Islamophobia, you name it.”
>
> “And these are quite apart from the obvious rape and death threats which say that ‘we will rape you in this fashion, you should be raped in this fashion, you should be killed in this fashion’, and so on and so forth.”
>
> – *Kavita Krishnan, Polit Bureau Member, Communist Party of India (ML) Liberation*

5. MUSLIM WOMEN RECEIVED 94.1% MORE ETHNIC OR RELIGIOUS SLURS THAN WOMEN FROM OTHER RELIGIONS

Muslim politicians (n = 7, 20.8% of problematic or abusive tweets) experienced more abuse than other groups, including women who are or perceived as Hindu (n = 66, 12.8% problematic or abusive content). Thus, Muslim women got 55.5% more problematic or abusive content when compared to other religions.

Frequency of problematic and abusive content by religion

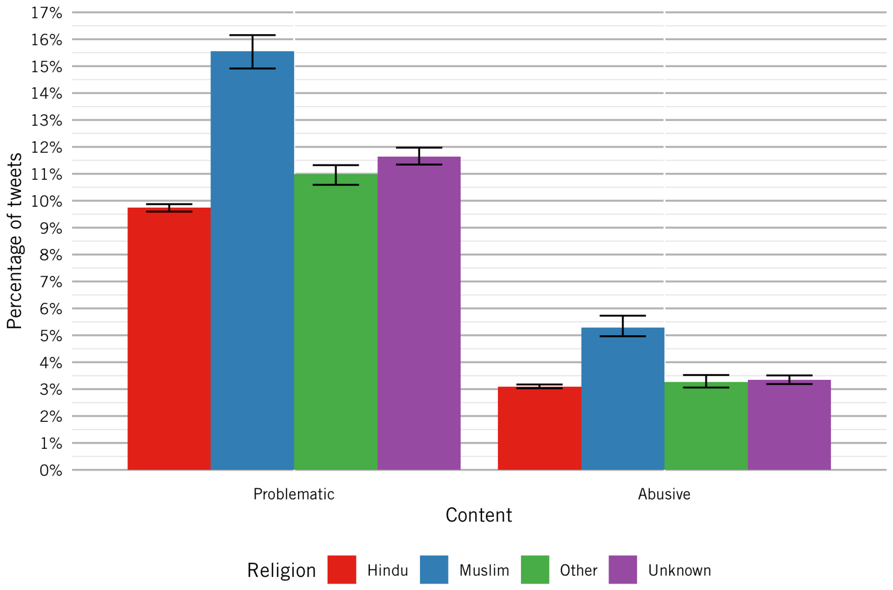


Fig: 8

Religion	Problematic	Abusive
Hindu	9.7%	3.1%
Muslim	15.6%	5.3%
Other	11.0%	3.3%
Unknown	11.6%	3.3%

In terms of type of abuse, Muslim women received 94.1% more ethnic or religious slurs than women from other religions. Racism-based abuse was also higher for muslim women at 12.6% vs 9.2% for Hindu women.

Frequency of type of problematic or abusive content by religion

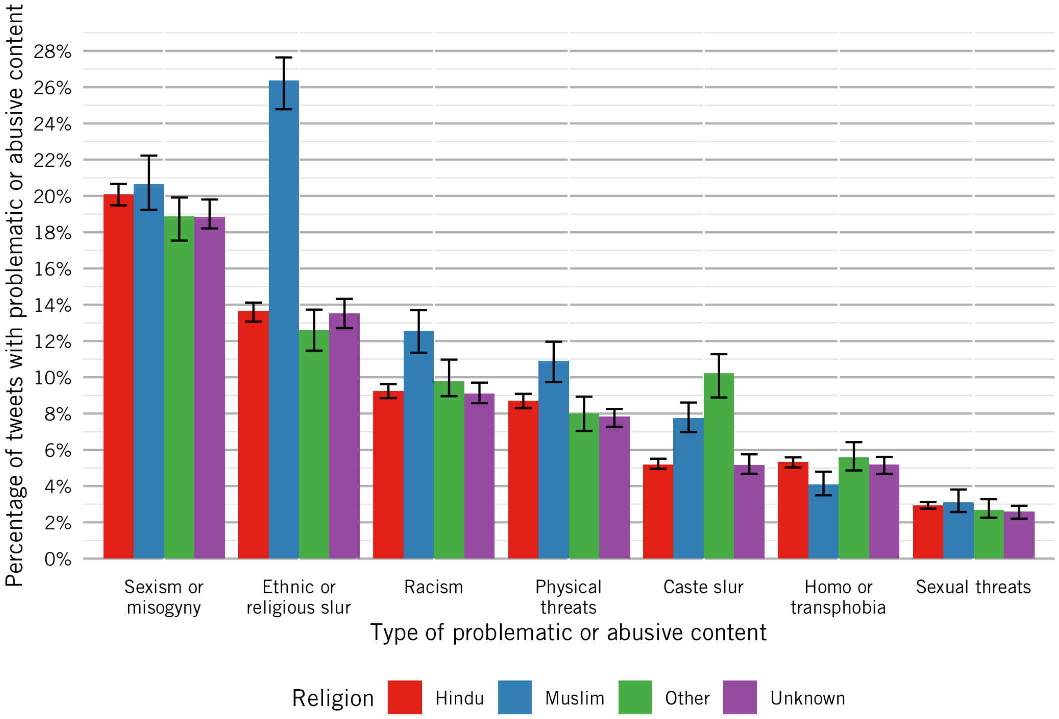


Fig: 9

Religion	Sexism or misogyny	Ethnic or religious slurs	Racism	Physical threats	Caste slurs	Homophobia or transphobia	Sexual threats
Hindu	20.1%	13.7%	9.2%	8.7%	5.2%	5.3%	2.9%
Muslim	20.7%	26.4%	12.6%	10.9%	7.7%	4.1%	3.1%
Other	18.9%	12.6%	9.8%	8.0%	10.2%	5.6%	2.7%
Unknown	18.9%	13.5%	9.1%	7.8%	5.2%	5.2%	2.6%

Comparison in the type of problematic and abusive content received by Muslim vs. non-Muslim women (ratio)

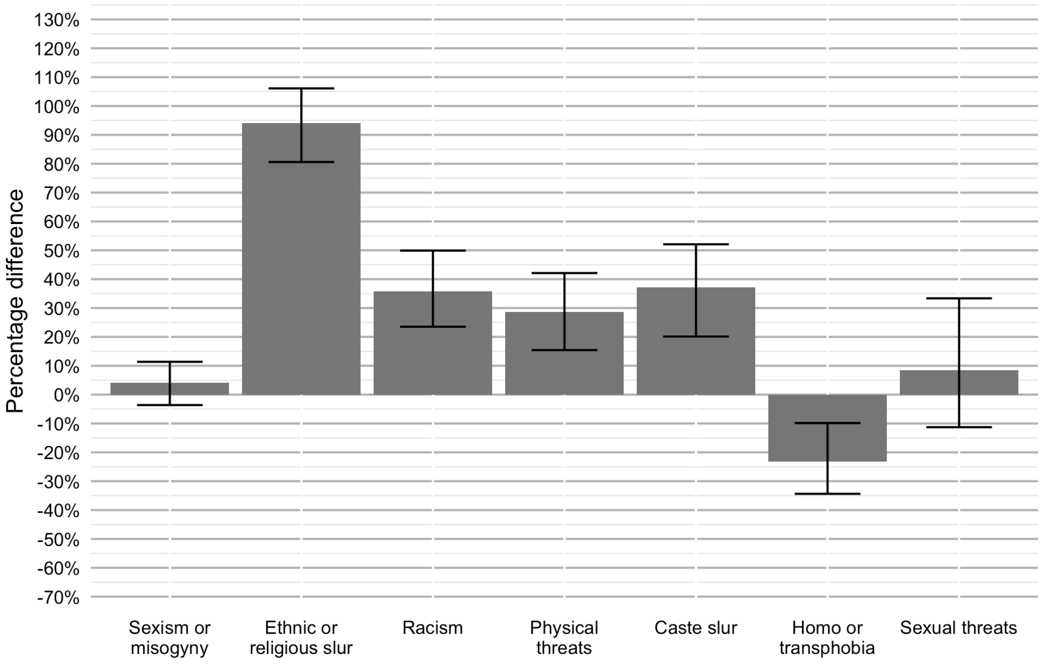


Fig: 10

Sexism or misogyny	Ethnic or religious slur	Racism	Physical threats	Caste slur	Homophobia or transphobia	Sexual threats
4.1%	94.1%	35.7%	28.7%	37.1%	-23.3%	8.5%

> “Being a Muslim woman sometimes becomes a huge burden. I am subjected to so much hate than a Muslim man. Only 25% of what I get is based on the content of my politics, 75-80 % is about being a woman and a Muslim woman.”
>
> *- Shazia Ilmi, Member, Bharatiya Janta Party*

6. WOMEN POLITICIANS BELONGING TO MARGINALISED CASTES RECEIVED 59% MORE CASTE-BASED ABUSE COMPARED TO WOMEN FROM OTHER CASTES

Women from marginalised castes received 59% more casteist slurs than women from general castes. In cases, where problematic or abusive content was identified, women belonging to marginalised castes (8.6%) received more caste-based slurs than those belonging to general (5.4%) or unknown/ undeclared castes (7.2%). This indicates that caste identity is more often than not, a key element of problematic or abusive content for women belonging to marginalised castes.

Frequency of type of abusive or problematic content by Caste

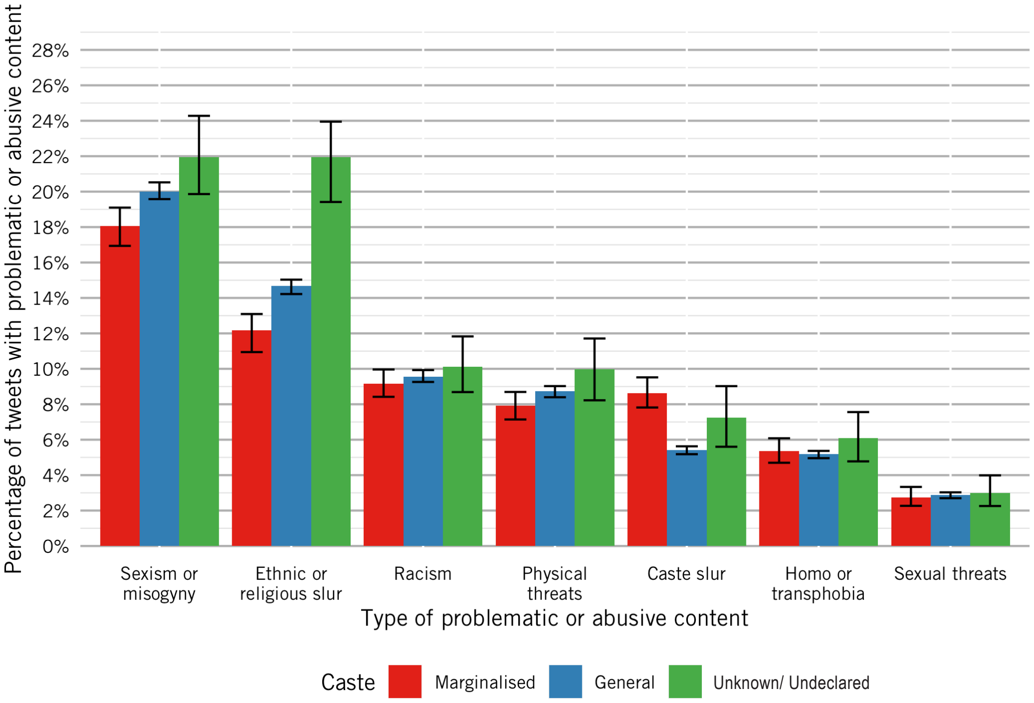


Fig: 11

Caste	Sexism or misogyny	Ethnic or religious slurs	Racism	Physical threats	Caste slurs	Homophobia or transphobia	Sexual threats
Marginalised	18.1%	12.2%	9.2%	7.9%	8.6%	5.4%	2.7%
General	20.0%	14.7%	9.6%	8.7%	5.4%	5.2%	2.9%
Unknown/ Undeclared	22.0%	21.9%	10.1%	10.0%	7.2%	6.1%	3.0%

Comparison in the type of problematic and abusive content received by women from marginalised castes vs. general castes (ratio)

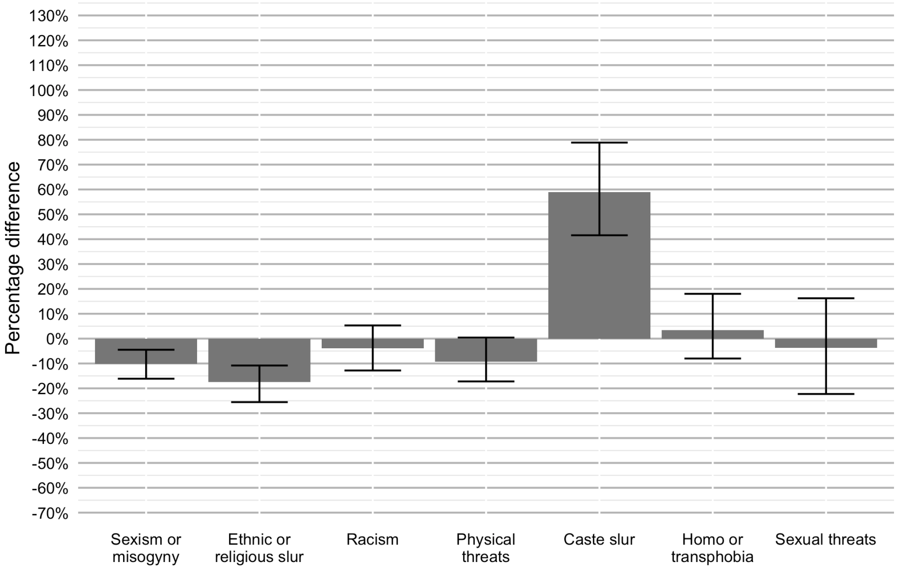


Fig: 12

Sexism or misogyny	Sexism or misogyny	Ethnic or religious slurs	Racism	Physical threats	Caste slurs	Homophobia or transphobia
-10.2%	-17.5%	-3.9%	-9.4%	59%	3.4%	-3.7%

@[redacted] अब तुझे अपनी जात और औकात का हसास जनता करा देगी
@[redacted] Tatti ka bacteria bola re😂😂😂
@[redacted] @[redacted] हम तो कुत्ता पालें हूं लेकिन ब्रिकिट ाकि टट्टी कराने के लिए क जरूरत है कोई बेरोज नौ मिल जाएगी
@[redacted] Tu reservation ki roh vahi layak hai tu chutiye
@[redacted] chup reh chamaaran

7. WOMEN POLITICIANS FROM POLITICAL PARTIES OTHER THAN THE BHARATIYA JANATA PARTY EXPERIENCED MORE ABUSE

Most of the politicians with tweet mentions (about 76%) belonged to either the Bharatiya Janata Party (BJP) or the Indian National Congress (INC), and about 62% of tweets in our sample mentioned someone from one of these parties. Compared to the ruling party BJP, women politicians from 'other parties' experienced 56.7% more problematic or abusive content than BJP while INC politicians received 45.3% more abusive or problematic content than BJP.

The 'other' parties include Aam Aadmi Party, Apna Dal, All India Anna Dravida Munnetra Kazhagam, All India Trinamool Congress, Bahujan Samaj Party, Community Party of India (Marxist), Communist Party of India (Marxist-Leninist) Liberation, Dravida Munnetra Kazhagam, Jammu & Kashmir National Congress, Jammu & Kashmir Peoples Democratic Party, Jharkhand Mukti Morcha, Rashtriya Janata Dal, Shiromani Akali Dal, Shiv Sena, Samajwadi Party, Telangana Rashtra Samiti, Yuvajana Sramika Rythu Congress Party. We have observed that, except for some notable cases, many women from these parties do not have an active Twitter account.

Frequency of problematic and abusive content received by women by political party

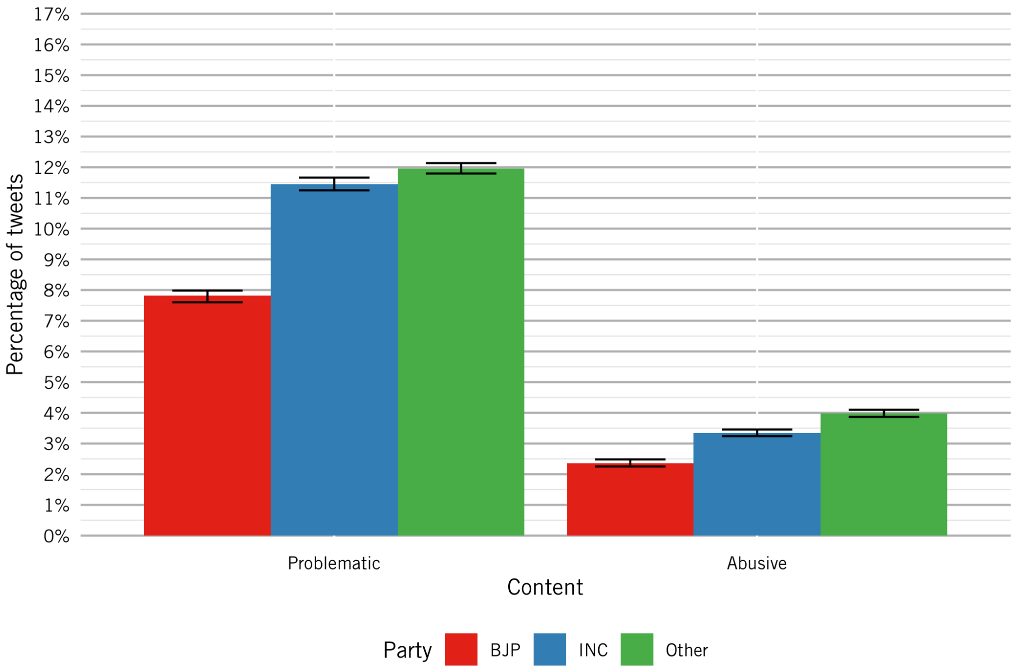


Fig: 13

Party	Problematic	Abusive
BJP	7.8%	2.4%
INC	11.5%	3.4%
Other	12.0%	4.0%

> "Most of the abuse is directed at Miss (Mehbooba) Mufti's twitter handle which I have taken over since September (2019). Previously the vile threats and abuses were targeting Ms. Mufti. Since the troll army knows it's me, it's been redirected towards me. Some of the threats I have received go into graphic detail about the manner in which I should be sexually assaulted and subsequently killed."
>
> – *Iltija Mufti, daughter of Mehbooba Mufti, the former Chief Minister of Jammu and Kashmir who is currently under administrative detention*

Looking into the type of abuse, showed that women from BJP received more ethnic or religious slurs (16.5%) compared to those from INC (11.6%).

Type of problematic or abusive content received by women by political party

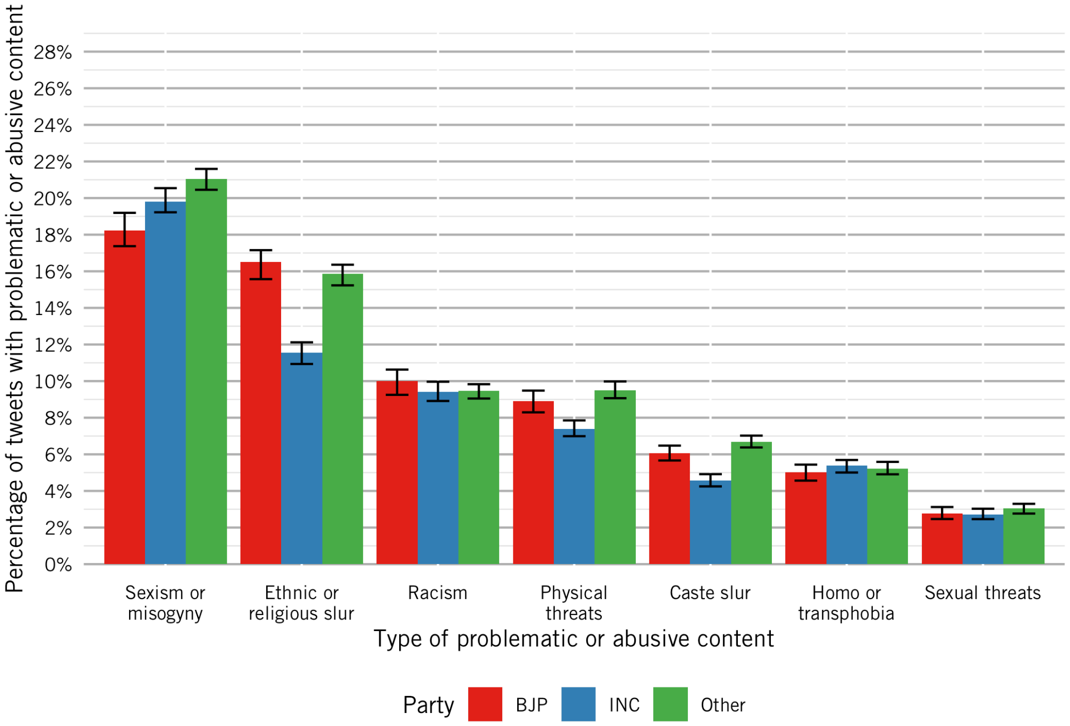


Fig: 14

Party	Sexism or misogyny	Ethnic or religious slur	Racism	Physical threats	Caste slur	Homophobia or transphobia	Sexual threats
BJP	18.2%	16.5%	10.0%	8.9%	6.1%	5.0%	2.8%
INC	19.8%	11.6%	9.4%	7.4%	4.6%	5.4%	2.7%
Other	21.1%	15.9%	9.5%	9.5%	6.7%	5.2%	3.1%

8. POLITICIANS WHO LOST THE ELECTION WERE TARGETED MORE

We also investigated the type of abuse by election results and noticed that politicians who lost, received 24.6% more problematic or abusive content than politicians who won the elections. They also received 59.3% more ethnic and religious slurs.

Frequency of problematic and abusive content by election outcome

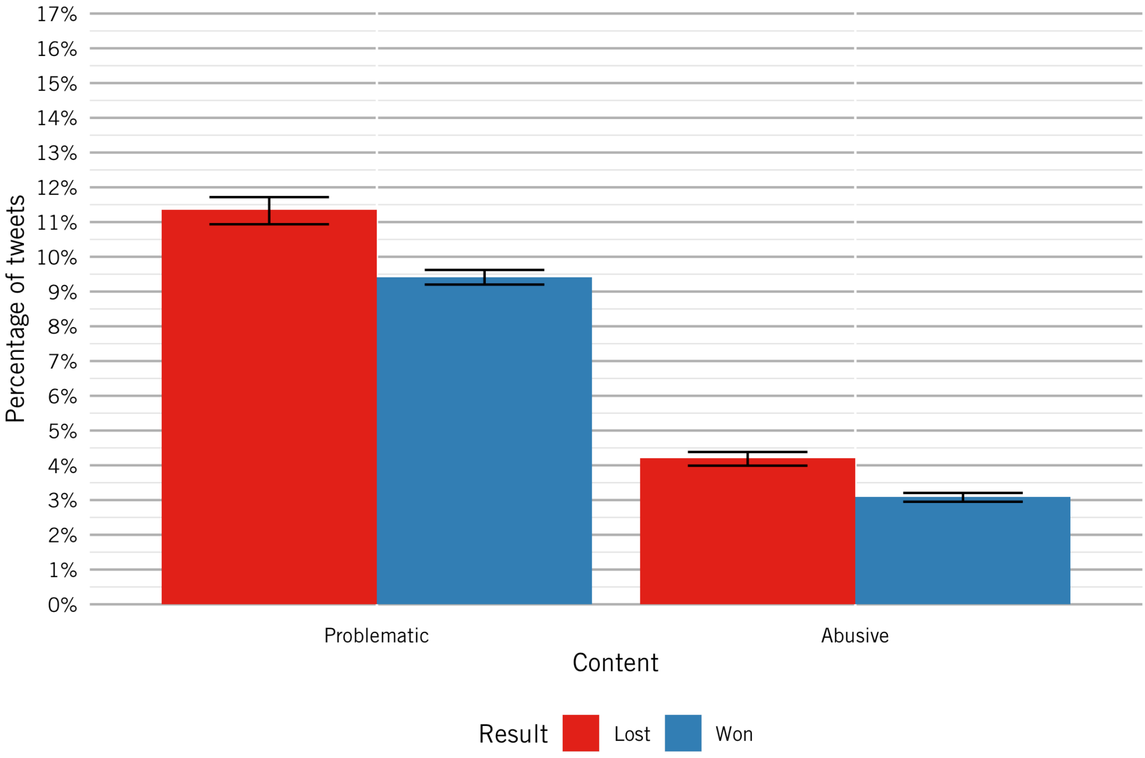


Fig: 15

Result	Problematic	Abusive
Lost	11.4%	4.2%
Won	9.4%	3.1%

Type of problematic or abusive content by election outcome

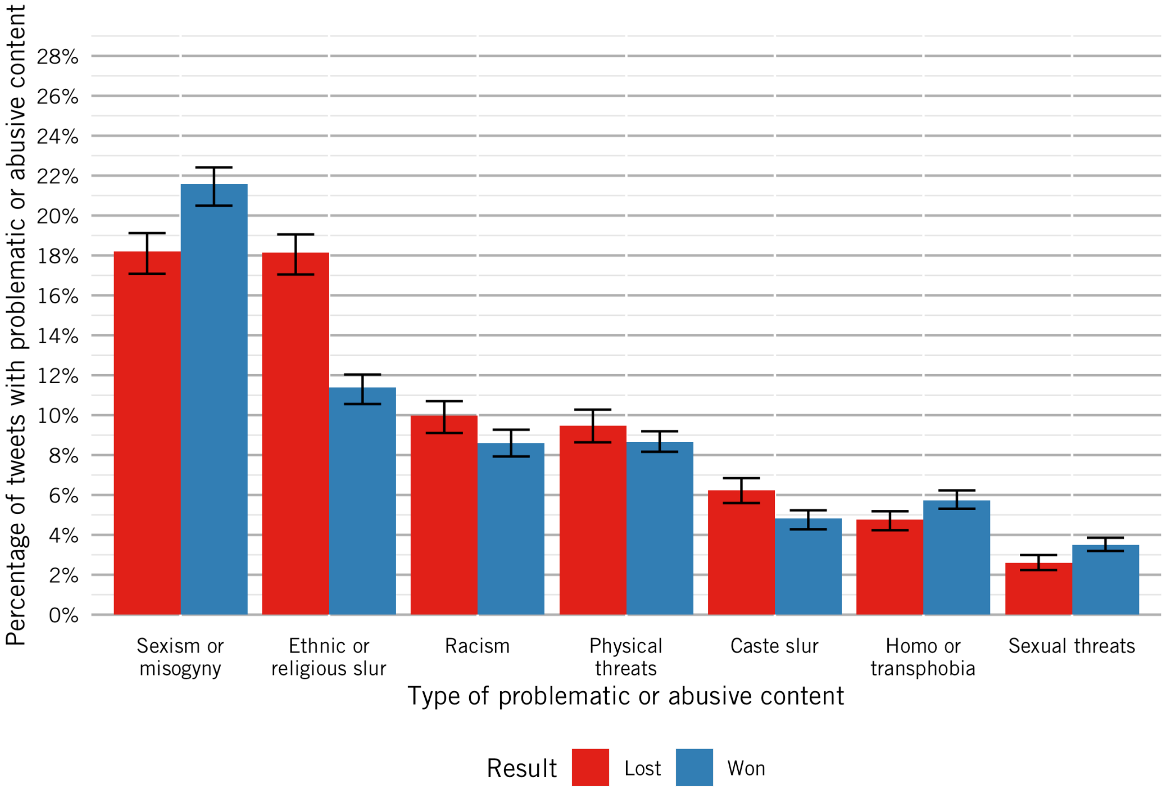


Fig: 16

Result	Sexism or misogyny	Ethnic or religious slurs	Racism	Physical threats	Caste slur	Homophobia or transphobia	Sexual threats
Lost	18.2%	18.1%	10.0%	9.5%	6.2%	4.8%	2.6%
Won	21.6%	11.4%	8.6%	8.7%	4.8%	5.7%	3.5%

9. UNMARRIED WOMEN WERE TARGETED MORE

We hypothesised based on the content of some tweets that there may be a difference between problematic and abusive content for women who are married as compared to those who are not. We found that politicians who were not currently married (including widowed, divorced, separated and unmarried) received 40.6% more abusive tweets and 31% more problematic tweets than married women. This suggests that women who are unmarried may be seen more as targets. We also noted that information on marital status of many politicians (33 of 95) was not publicly available.

Frequency of abusive and problematic content by marital status

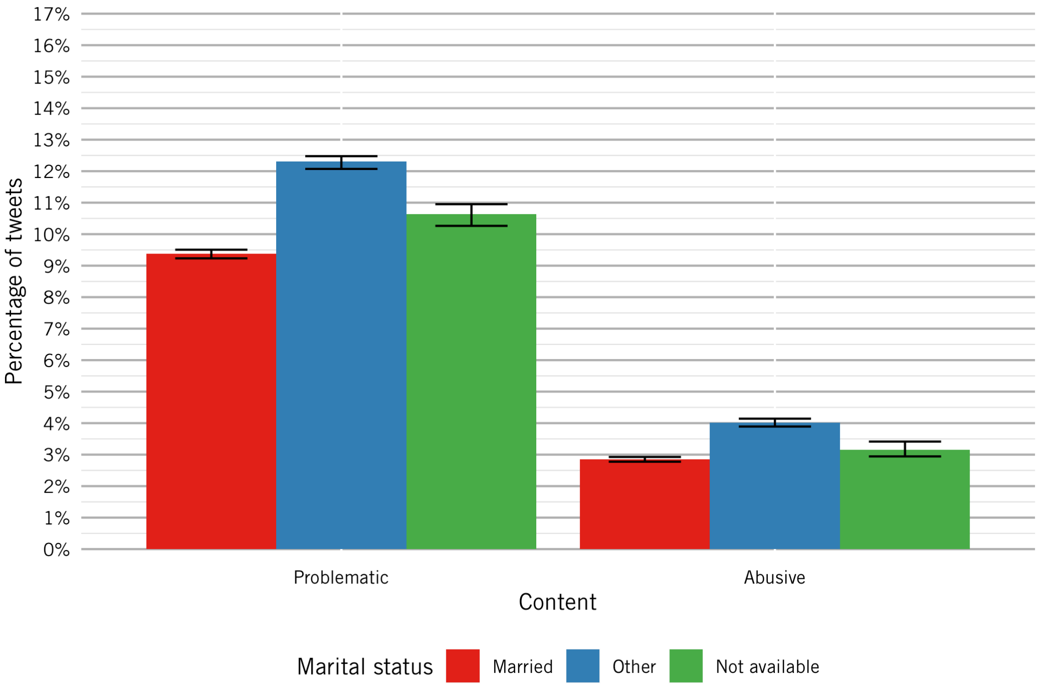


Fig: 17

Marital status	Problematic	Abusive
Married	9.4%	2.9%
Other	12.3%	4.0%
Not available	10.6%	3.2%

The most frequent type of abuse received by unmarried women was 'sexism or misogyny', 13.9% more frequent than for married women.

Frequency of type of problematic or abusive content by marital status

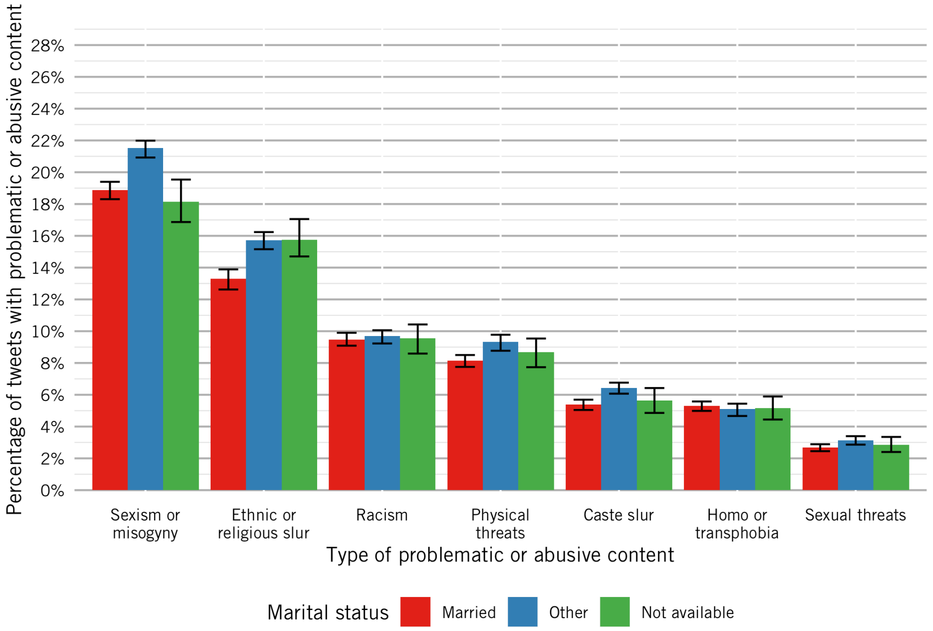


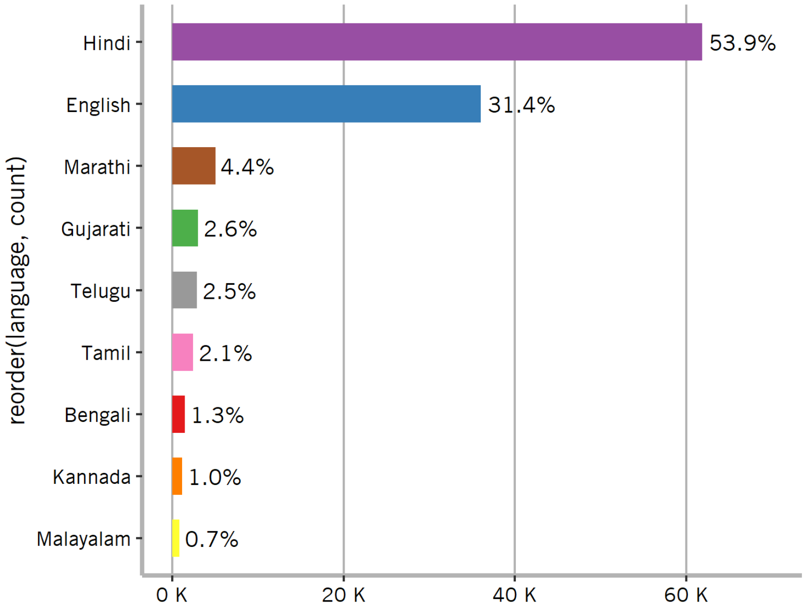
Fig: 18

Marital status	Sexism or misogyny	Ethnic or religious slurs	Racism	Physical threats	Caste slurs	Homophobia or transphobia	Sexual threats
Married	18.9%	13.3%	9.5%	8.1%	5.4%	5.3%	2.7%
Other	21.5%	15.7%	9.7%	9.3%	6.4%	5.1%	3.1%
Not available	18.1%	15.8%	9.6%	8.7%	5.6%	5.2%	2.8%

10. HINDI LANGUAGE WAS MOST FREQUENTLY USED TO ABUSE WOMEN POLITICIANS

We looked at overall rates of problematic and abusive content among the different detected languages. The language distribution as corrected for sampling is shown below. As we have noted in Annexure 2, these were detected languages using an automatic language detection tool. The language detection had errors. These languages were used in the Decoders platforms, allowing the Decoders to select preferred languages before labelling tweets. They could select one or multiple languages and they were only shown tweets in their preferred languages.

Languages of Tweets



Language	Proportion
Hindi	53.9%
English	31.4%
Marathi	4.4%
Gujarati	2.6%
Telugu	2.5%
Tamil	2.1%
Bengali	1.3%
Kannada	1.10%
Malayalam	0.7%

Fig: 19

The language with most problematic or abusive content was also the most common - Hindi (15.3% all problematic or abusive content). We found that problematic or abusive content was 26.9% more frequent in Hindi than other languages in our study (with 27.7% more problematic content and 24.6% more abusive content).

All languages showed the expected proportions of problematic and abusive content except for Tamil. We investigated this anomaly more closely and we believe it's largely due to one "superuser" Decoder who had labelled more tweets as abusive than problematic. This was not consistent with other Decoders or with two experts (Amnesty staff) who labelled 300 tweets as a test, but the overall proportion of problematic and abusive tweets was still consistent with other languages and with Amnesty's internal experts, and so this user's contributions were still helpful in generating our findings.

Frequency of problematic or abusive content by tweet language

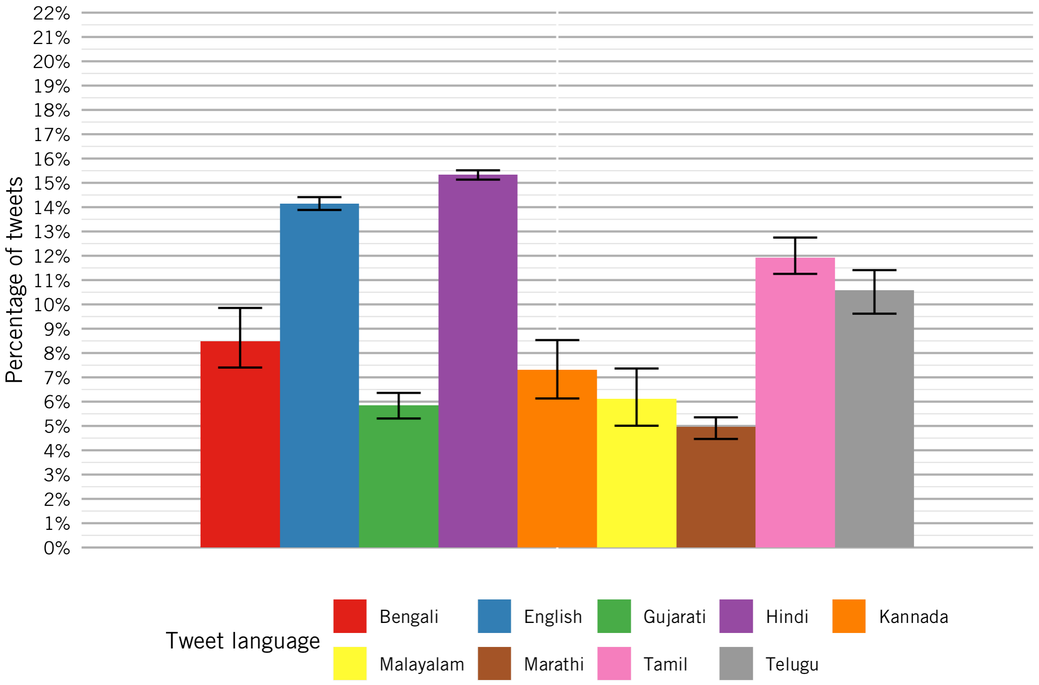


Fig: 20

Notes:

- *The black bars represent confidence intervals. There was high confidence for languages with a large volume of tweets in the sample (such as English and Hindi) and a much lower confidence for languages like Telugu, Malayalam or Kannada.*

Tweet language	Problematic or abusive
Bengali	8.5%
English	14.1%
Gujarati	5.8%
Hindi	15.3%
Kannada	7.3%
Malayalam	6.1%
Marathi	5.0%
Tamil	11.9%
Telugu	10.6%

11. THERE WAS A DECREASE IN ABUSE IN ENGLISH LANGUAGE BUT NOT IN HINDI LANGUAGE DURING ELECTION

We noted that the proportion of abusive tweets as compared to non-abusive ones was lower during the election dates than in the month preceding the elections.

Proportion of abusive, problematic and not abusive tweets grouped by period

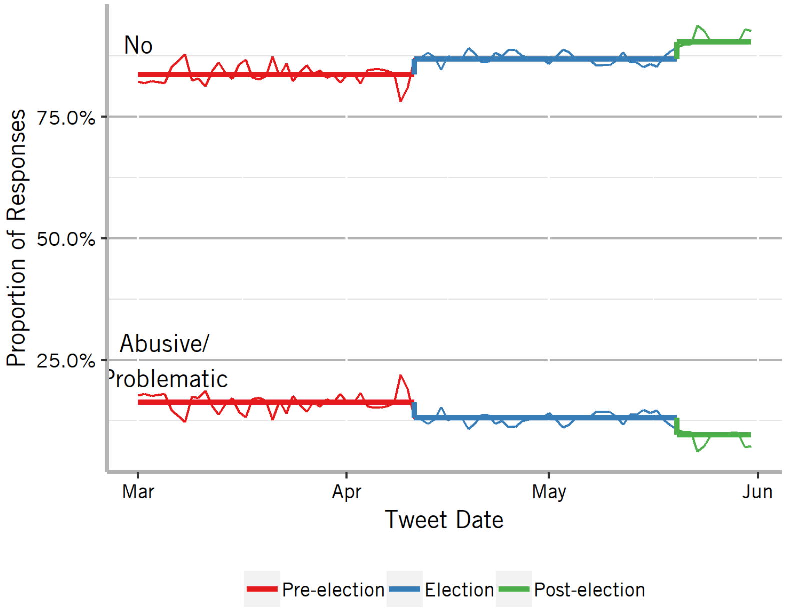


Fig: 21

Frequency of problematic and abusive content by period

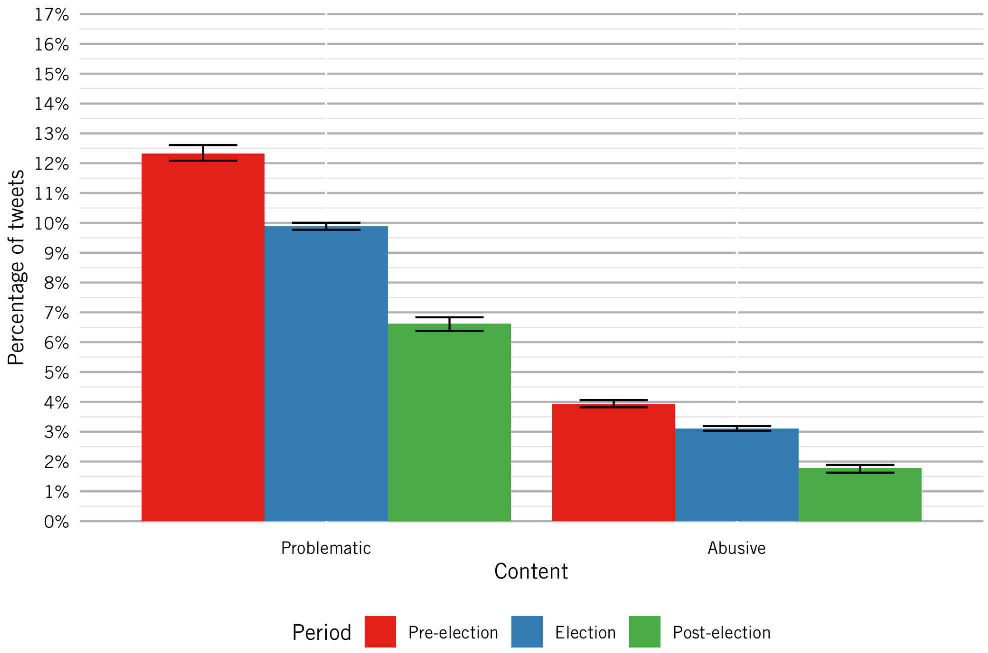


Fig: 22

Period	Problematic	Abusive
Pre-election	12.3%	3.9%
Election	9.9%	3.1%
Post-election	6.6%	1.8%

Type of problematic or abusive content by period

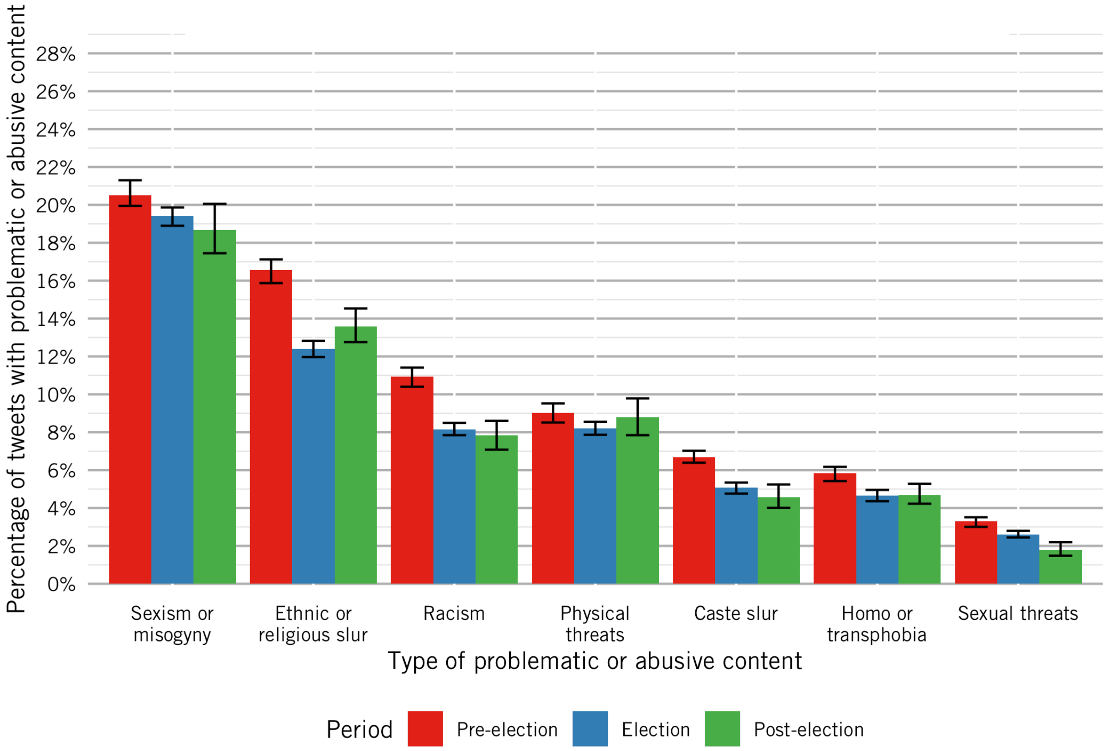


Fig: 23

Period	Sexism or misogyny	Ethnic or religious slurs	Racism	Physical threats	Caste slurs	Homophobia or transphobia	Sexual threats
Pre-election	20.5%	16.6%	10.9%	9.0%	6.7%	5.8%	3.3%
Election	19.4%	12.4%	8.1%	8.2%	5.1%	4.7%	2.6%
Post-election	18.7%	13.6%	7.8%	8.8%	4.6%	4.7%	1.8%

To understand this phenomena, Amnesty International India corresponded with Twitter on 18 November 2019 asking the following besides other questions:[26]

> *“Specifically, please list any measures taken by Twitter to decrease levels of online abuse during the 2019 Indian General Elections, such as but not limited to increased monitoring, increased numbers of moderators, increased use of automated process to detect violent and abusive content, de-activation of accounts, etc. Please describe the measures taken and any evaluation conducted to understand the effects of such measures.”*

Twitter, in its response dated 29 November 2019, has detailed measures taken in the period of January to June 2019. The response states,[27]

> *“Across Twitter, more than 50% of Tweets we took action on for abuse were proactively surfaced using technology, rather than relying on reports from people who use Twitter.”*

> *“Over this period we saw a 105% increase in accounts actioned by Twitter (locked or suspended for violating the Twitter Rules).”*

> *“There was a 48% increase in accounts reported for potential violations of our Hateful Conduct policies. We actioned 133% more accounts compared to the last reporting period. Similarly, we saw a 22% increase in accounts reported for potential violations of our abuse policies. We took action on 68% more accounts compared to the last reporting period.”*

Amnesty International India noted that, it was unclear if any specific measures adopted by Twitter, or some external factors, contributed to the drop in frequency of abuse. So as to understand this better, we examined this drop in abuse in terms of language.

[26] *See*, Annexure 6, Letter to Twitter

[27] *See*, Annexure 7, Twitter's Response

DROP IN ABUSE IN ENGLISH LANGUAGE WHILE ABUSE IN HINDI LANGUAGE REMAINED CONSTANT

We wanted to check if all languages showed this pattern of drop in abuse during Election dates. Looking more closely, it appeared that most of the drop in abuse proportion was in English, while Hindi remained constant. We do not have an explanation for this drop, but it could be hypothesised that certain measures by Twitter could have reduced abuse in English. If this hypothesis be true, it highlights the importance of recognising the nuances of language in detection of online abuse.

Problematic or abusive tweets by period in English and Hindi

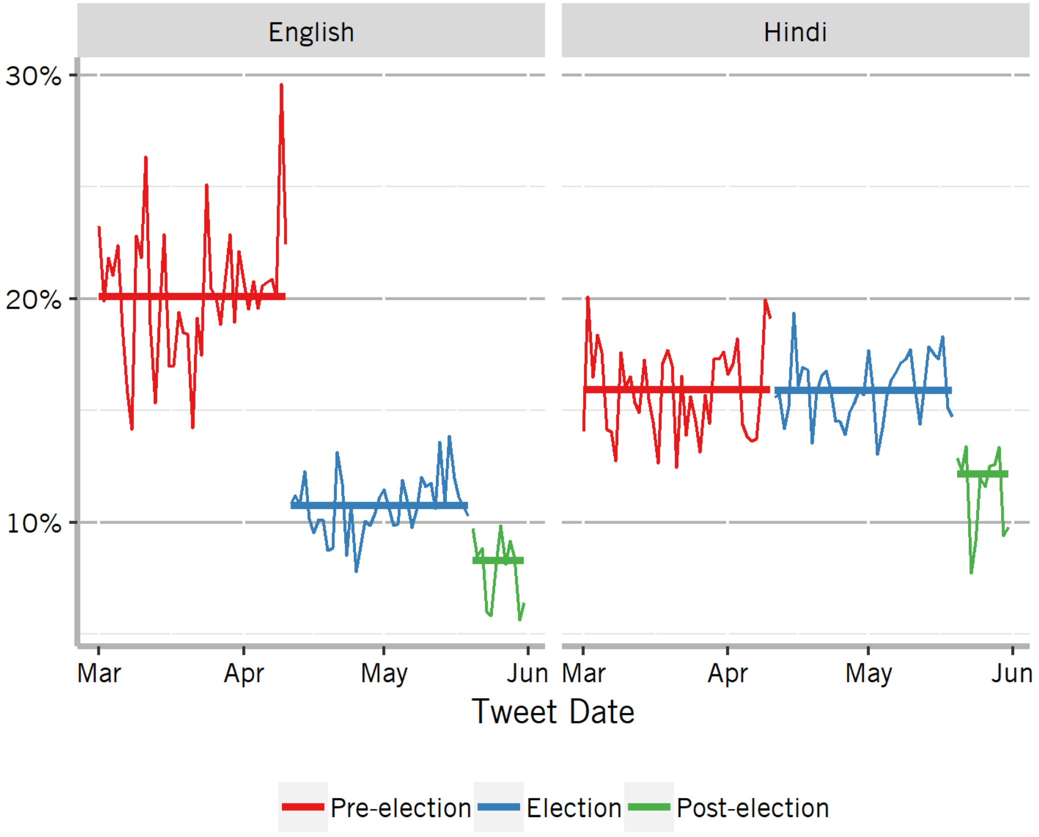


Fig: 24

Period	Problematic or abusive in English	Problematic or abusive in Hindi
Pre-election	20.3%	15.9%
Election	10.8%	16.0%
Post-election	7.5%	10.7%

Twitter: Human Rights Responsibilities

Companies, wherever they operate in the world, have a responsibility to respect all human rights. This is an internationally endorsed standard of expected conduct.[28]

[28] Guiding Principles on Business and Human Rights – Implementing the United Nations "Protect, Respect and Remedy" Framework, United Nations Human Rights Office of the High Commissioner, 2011, https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf

INTERNATIONAL LAW STANDARDS

International human rights standards classify violence against women as a form of discrimination that requires comprehensive responses.[29] Online violence, as a form of violence against women, "extends to any act of gender-based violence against women that is committed, assisted or aggravated in part or fully by the use of Information Communications and Technology (ICT), such as mobile phones and smartphones, the Internet, social media platforms or email, against a woman because she is a woman, or affects women disproportionately."[30]

In 2018, the Special Rapporteur on online violence against women said that "women are both disproportionately targeted by online violence and suffer disproportionately serious consequences as a result'.[31]

Similar to offline spaces, the discrimination faced by women online is intersectional and affected by many other factors such as race, ethnicity, caste, sexual orientation, gender identity and expression, abilities, age, class, income, culture, religion, and urban or rural setting.

To address this discrimination, the United Nations Human Rights Council affirmed that the rights women hold in the offline realm must be guaranteed to them in online spaces too.[32] This includes a woman's right not to be subject to gender-based violence, the right to freedom of expression, the right to privacy and the right to have access to information shared through ICT.[33] In reality, these rights are being constantly violated.[34] Online violence results in multiple layers of marginalisation, as abusive content is disseminated and shared by others, further perpetuating such violence.

TWITTER'S RESPONSE

In 2018, responding to Amnesty International, Twitter committed that that they are "energized and motivated" to address abuse and hateful conduct directed at women.[35] In November 2019, Amnesty International India reached out to Twitter for its response on specific questions emanating from this study.[36]

In its response to Amnesty International India in November 2019, Twitter reaffirmed that "building Twitter, free of abuse, spam and other behaviour that distract from the public conversation is one of their top priorities".[37] It further shared that it has "made strides in creating a healthier service….. to positively and directly impact people's experience on the service".

Amnesty International India and Amnesty International acknowledge that in the last few years, Twitter has taken positive steps towards addressing the problem of violence and abuse against women on the platform. Some of these include enhancing features for reporting abuse, educating users - primarily women, teachers and non-governmental organisations on online safety, reviewing and updating various rules, among others.

However, as highlighted by the Findings of this study, the scale and nature of online abuse that women face is significantly high. Women have the right to live free from discrimination and violence. They also have the right to freely express themselves, both online and offline. Twitter needs to further strengthen and enhance its policies and rules, particularly in diverse cultural contexts, so as to adequately meet its responsibility towards women engaging on the platform. As admitted by Twitter CEO Jack Dorsey, Twitter has created a "pretty terrible situation" for women, and it is "super easy" to harass on Twitter.[38]

TWITTER'S POLICIES AND PROCESSES ARE NOT TRANSPARENT

The UN Guiding Principles on Business and Human Rights[39] states that the responsibility to respect human rights does not only mean having policies and processes that respect human rights but also 'showing' commitment towards respect for human rights in practice.[40] As the guiding principles explain, "showing involves communication, providing a measure of transparency and accountability to individuals and groups who may be impacted".[41]

Twitter touts transparency as being essential to its human rights responsibilities.[42] Accordingly, Twitter's reporting mechanisms should be accessible and transparent.[43]

As actions for violation of their policy, Twitter lays down a range of enforcement options. In its 2018 Transparency Report, Twitter said, "Context matters when evaluating reports of abusive behavior and determining appropriate enforcement actions".[44] However, the criteria to assess the severity of violence and appropriate resolutions is not provided.[45] For example, one enforcement option is to place the account in read-only mode 'If it seems [to Twitter] like an otherwise healthy account is in the middle of an abusive episode'.[46] The access is restored when "calmer heads prevail."[47] Such conditions are subjective and open-ended, and lack of further explanations grant wide discretionary powers to Twitter, at the cost of users, who are left uncertain as to their recourse in the face of abusive behaviour targeting them on the platform.

Twitter also states that users can appeal a decision based on their review of a report of violence and abuse if the user believes that they made an error.[48] A detailed overview of the appeal process, including an explicit commitment to respond to all appeals or a timeframe of when to expect a response is not included in any of Twitter's policies.[49]

As stated by Twitter itself, they have received feedback that their policies need to be more precise and clear. Additionally, their policies need to constantly evolve in order to protect marginalised population.[50] In July 2019, Twitter refreshed its rules to include "simple, clear language".[51] However, it lacked focus on the safety of women and other marginalised populations.

In terms of training provided to its content moderators, Twitter shared with Amnesty International India, that the "team undergoes in-depth training on their policies, ensuring that social and political nuances, local context and cultures is taken into account." Their employee assistance programs is mindful of the work of moderators which requires them to review sensitive content.[52]

We acknowledge Twitter's response. However, given the scale of the abuse found by this study, it is important that Twitter is more transparent with the details regarding the trainings it provides to its moderators, including specialised training on gender, regional context, languages etc. Information should be made available on the number of moderators, the volume of tweets evaluated by them, the time taken, and the checks and balance put in place to ensure their impartiality and efficiency. Twitter must also ensure that moderators' rights are respected in this process, including their rights to the highest attainable physical and mental health in the workplace. It must also ensure that they do not suffer the adverse impacts from repeated or sustained exposure to traumatic content without adequate support and training.[53] Such transparency will allow external appraisal of the efficacy of Twitter's human rights efforts and whether the policies are being applied impartially and with caution.

[29] Convention on Elimination of All Forms of Discrimination against Women

[30] Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective on 14 June 2018, A/HRC/38/47

[31] *Ibid*

[32] UN Human Rights Council. The promotion, protection and enjoyment of human rights on the Internet: resolution adopted by the Human Rights Council on 1 July 2016. A/HRC/RES/32/13

[33] Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective on 14 June 2018, A/HRC/38/47

[34] Toxic Twitter – A Toxic Place for Women, Amnesty International, https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1/

[35] Letter from Twitter to Amnesty International, 15 Mar. 2018, https://amnesty.app.box.com/s/j4z4h3qlkfri0kwf7dpbvrpv2i5zgye3

[36] For letter from Amnesty International India to Twitter India, 18 November 2019 *see,* Annexure 6.

[37] For letter from Twitter India to Amnesty International India, 29 November 2019 *see,* Annexure 7.

[38] Aria Bendix, Jack Dorsey Says Twitter Makes It 'Super Easy' to Harass and Abuse Others, Entrepreneur India, 17 Apr. 2019, https://www.entrepreneur.com/article/332408

[39] Guiding Principles on Business and Human Rights, https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf

[40] Amnesty International, Toxic Twitter, Chapter 7, INDEX NO. ACT 30/8070/2018

[41] Guiding Principles on Business and Human Rights, https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf

[42] Twitter Transparency Report, https://transparency.twitter.com/

[43] Zarizana Abdul Aziz, Due Diligence and Accountability for Online Violence Against Women, Association for Progressive Communications, July 2017 https://www.apc.org/sites/default/files/DueDiligenceAndAccountabilityForOnlineVAW.pdf

[44] Twitter Transparency Report, Twitter Rules Enforcement, Abuse Policy Enforcement: https://transparency.twitter.com/en/twitter-rules-enforcement.html

[45] Amnesty International, Toxic Twitter, Chapter 7, INDEX NO. ACT 30/8070/2018

[46] "It limits their ability to Tweet, Retweet, or Like Content", Our Range of Enforcement Options, https://help.twitter.com/en/rules-and-policies/enforcement-options

[47] Our Range of Enforcement Options, https://help.twitter.com/en/rules-and-policies/enforcement-options

[48] *Ibid*

[49] Amnesty International, Toxic Twitter, Chapter 4, INDEX NO. ACT 30/8070/2018

[50] https://blog.Twitter.com/official/en_us/topics/company/2019/hatefulconductupdate.html

[51] Letter from Twitter India to Amnesty International India, 29 Nov 2019; See also, Del Harvey, Making Our Rules Easier to Understand, 7 Jun. 2019, https://blog.Twitter.com/en_us/topics/company/2019/rules-refresh.html

[52] Letter to Amnesty International India from Twitter, 29 November 2019, *see* Annexure 7.

[53] *See* UN Committee on Economic, Social and Cultural Rights, General Comment 18 (Right to Work), para. 6, "Work as specified in article 6 of the Covenant must be decent work. This is work that respects the fundamental rights of the human person as well as the rights of workers in terms of conditions of work safety and remuneration. It also provides an income allowing workers to support themselves and their families as highlighted in article 7 of the Covenant. These fundamental rights also include respect for the physical and mental integrity of the worker in the exercise of his/her employment." [emphasis added].

In terms of Amnesty International India's specific question on Twitter's internal mechanism for using language detection tools and machine learning for Reporting (abuse) and Enforcements, Twitter shared that they "use a combination of human review and technology to help them enforce their rules. Their team reviews and responds to reports, 24/7; and they have the capacity to review and respond to reports in multiple languages."[54]

This study has found that during the election dates, there was a significant decrease in abuse in English, while abuse in Hindi remained constant. As discussed in the Findings chapter, Amnesty International India hypothesises that this anomaly could be because of certain tools and measures adopted by Twitter during election dates in India. While the results was positive for English language, if this hypothesis be true, this highlights the importance of recognising the nuances of language in detection of online abuse.

In addition to this, Twitter needs to ensure that the use of automated tools should only take place where there is a "human in the loop" and should form part of a larger content moderation system characterised by human judgement, greater transparency, right to appeal and other safeguards. Automated systems that are used as the sole mechanism to take down content poses a serious risk of restricting legitimate expression online.[55]

TWITTER NEEDS TO ENABLE AND EMPOWER USERS

Enabling and empowering users to experience a safe Twitter experience is a key component of Twitter's human rights responsibility.

Twitter has shared with Amnesty International India that the platform provides "a series of tools to help people keep safe and give them control over what they see and who they interact with".[56] These tools include features such as unfollow, block, advanced block, mute, disable receive direct message setting filtered notifications, protected tweets, safe search and sensitive media.

These privacy features, while enabling users to utilise individual safety measures, could be further enhanced by taking feedback from users who have reported abuse – whether they were satisfied with the reporting process and the action taken.

In terms of Safety and Awareness campaigns, Twitter has shared with Amnesty International India that "it runs a number of public campaigns aimed at increasing awareness on online safety and helping people who use Twitter, take control of their online experience". Twitter has also engaged safety partners in India to help Twitter in its policies.

The safety and awareness campaigns initiated by Twitter are a positive step towards addressing online abuse. Twitter should expand this initiative further by focusing not just on women users and marginalised groups, but on users in general – women, men and non-binary people. The Twitter platform could constantly reach out to its users in creative ways to discourage them from engaging in abusive behaviour.

As affirmed by Colin Crowell, Global Vice President of Public Policy of Twitter:

"Not many Twitter users in India are aware of how to report abuse or harassment they face on the open communication platform, opting for the wrong way of posting an abusive photo or tweet and then requesting us to take action".[57]

By enhancing and strengthening the awareness campaigns on the reporting tools available and evaluating the effectiveness of the measures in place to effectively tackle online violence against women, Twitter can ensure that it is empowering and enabling women and all its users.

TWITTER FAILS TO PREVENT DISCRIMINATION, PARTICULARLY AGAINST WOMEN AND MARGINALISED COMMUNITIES

Twitter has a human rights responsibility to ensure that its policies respect users' right to free expression, without discrimination, especially on the basis of gender, religion, ethnicity and race.

Twitter must proactively take steps to affirm its commitment to protect users from discrimination, by putting in place mitigation measures. These mitigation measures should be effective in ensuring that women, minority communities, marginalised caste groups, non-binary users and other groups experiencing discrimination, feel safe and respected on the platform.

Twitter has introduced specialised policies that consider content based on child sexual exploitation[58] and terrorism and violent extremism,[59] to be violative of its norms. It should also enact a separate policy on enabling safe online spaces for women, reinforcing their commitment to gender equality and to the protection and empowerment of women.

Responding to Amnesty International India's specific question on Twitter's mechanisms related to content moderation and language, Twitter shared that "our focus is on ensuring we are covering the most widely used languages on Twitter in each market. Our global team undergoes in-depth training into our policies, and we also have an intensive focus on local language, culture, and context, ensuring we're taking social and political nuances into account. For example, we have native language speakers in major Indic languages used on Twitter".

Twitter's global rules are sometimes rendered ineffective by cultural differences. The abuse in India is often in a variety of regional languages, including colloquial slang or Hindi in English (latin) script. All this potentially escapes the radar of Twitter's auto language detection. Sexist expletives in Indian languages, with culture-specific meanings, are often used to abuse women. Twitter's content moderators may not be equipped to understand and label slang and local language as abuse. As mentioned in their update regarding the General Elections of India in 2019, Twitter has a global team dedicated to enforcing impartiality in elections, and the India team does not make enforcement decisions.[60]

Twitter needs to be proactive in ensuring that it is not engaging, whether deliberately or by consequence, in any discrimination, including discrimination on the basis of caste.

It is important that Twitter recognises the diverse cultural contexts and realities of its user countries and strengthens its policies to make it inclusive. By including caste discrimination as a disaggregated form of abuse, Twitter will commit to protect all marginalised groups in India, including Scheduled Castes/ Scheduled Tribes/ Other Backward Classes.

[54] Letter to Amnesty International India from Twitter, dated 29 November 2019, *see* Annexure 7.

[55] Amnesty International India's submission to the Ministry of Electronics and Information Technology, Government on the draft Information Technology (Intermediary Guidelines) Rules 2018, Amnesty International India, 30 Jan. 2019, https://amnesty.org.in/wp-content/uploads/2019/01/Amnesty-India-submission-on-IT-rules-30-Jan-2019-1.pdf

[56] Letter to Amnesty India from Twitter, dated 29 November 2019

[57] Most Indian Users Unaware How to Report Abuse: Twitter, Times Now News, 5 Mar. 2019, https://www.timesnownews.com/technology-science/article/most-indian-users-unawarehow-to-report-abuse-Twitter/376709

[58] Child Sexual Exploitation Policy, Twitter, https://help.twitter.com/en/rules-and-policies/sexual-exploitation-policy

[59] Terrorism and Violent Extremism Policy, Twitter, https://help.twitter.com/en/rules-and-policies/violent-groups

[60] Colin Crowell, Setting the record straight on Twitter India and impartiality, 8 Feb. 2019, Twitter, https://blog.twitter.com/en_in/topics/events/2019/impartiality.html

TWITTER TRANSPARENCY REPORT

Twitter publishes a Transparency Report biannually. In its latest report covering the first six months of 2019, Twitter claimed that it had received the highest number of global legal demands to remove content since the transparency report was launched in 2012 (67% more global legal demands to remove content).[61]

INFORMATION REQUESTS

Twitter reported that it received approximately 6% more global information requests. There were 79 emergency disclosure and 36 account preservation requests filed by the Indian Government. They also reported two account information requests from non-government entities in India.[62]

REMOVAL REQUESTS

Twitter reported that it withheld 42 accounts and 23 Tweets in India in 2019 so far, in response to 13 Blocking Orders from the Ministry of Electronics and Information Technology.[63] During the General Elections 2019, Twitter reported that it received 21 requests from the Election Commission of India and that it withheld 117 Tweets.[64]

TWITTER RULES ENFORCEMENT

Twitter reported a 42% increase in the number of unique accounts reported.[65] They also reported a 21% increase in the accounts reported by Government entities.[66]

[61] Transparency Report, 'Removal Requests', Twitter, https://transparency.twitter.com/en/removal-requests.html

[62] Transparency Report, 'Information Request', Twitter, https://transparency.twitter.com/en/information-requests.html

[63] The orders were issued under Section 69A of the Information Technology Act, 2000 for disseminating objectionable content in order to prevent incitement to harm or public disorder, Removal Requests, Twitter, https://transparency.twitter.com/en/removal-requests.html

[64] This was done under the Representation of the People Act, 1951 and the Indian Penal Code's relevant articles on election and election silence periods https://transparency.twitter.com/en/removal-requests.html

[65] These accounts were reported across the seven Twitter Rules policy categories of abuse, child sexual exploitation, hateful conduct, impersonation, private information, sensitive media and violent threats.

[66] Transparency Report, 'Twitter Rules Enforcement', Twitter, https://transparency.twitter.com/en/twitter-rules-enforcement.html

TWITTER GUIDELINES

Twitter has a number of policies that users have to abide by. Violations of these policies could result in enforcement actions and penalties. Apart from its Global policies, Twitter also updates its policies with regards to specific events such as the General Elections in India in 2019.

Twitter stated that **"India is the world's largest democracy, and one of our largest and fastest-growing audience markets in the world so the 2019 Lok Sabha is a key priority for Twitter, globally".**[67]

Twitter further stated that so as to "protect and enhance the health of the public conversation at this pivotal cultural and political moment", it has made a number of changes to its product, policies and approach to enforcement to address the behaviors which distort and detract from the public conversation on Twitter - particularly those which can surface at critical moments such as elections.[68]

HATEFUL CONDUCT POLICY - Promoting violence, inciting or wishing harm, attacking people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease is prohibited on Twitter.[69] In July 2019, Twitter updated its Hateful Conduct Policy to include "language that dehumanizes others on the basis of religion".[70] They also issued examples of what would constitute hateful conduct.[71]

ABUSIVE BEHAVIOR POLICY - Targeted harassment or incitement of other people to do so is banned. Abusive behavior is defined to include unwanted sexual advances and sexually objectifying content.[72]

VIOLENT THREATS POLICY - Threatening violence is prohibited. This includes threatening to sexually assault someone.[73]

GLORIFICATION OF VIOLENCE POLICY - The glorification of violence is also banned.[74]

PRIVATE INFORMATION POLICY - Publishing or threatening to publish private information without authorization or incentivizing others to do so is prohibited.[75]

SENSITIVE MEDIA POLICY - Media depicting sexual violence, assault, graphic violence, adult content and hateful imagery are prohibited under this policy.[76]

Apart from these policies, Twitter also prohibits using the platform for the promotion of self-harm or suicide, child exploitation, terrorism or violent extremism, illegal activities, non-consensual nudity, platform manipulation and spam, manipulating or interfering in elections, impersonation and the violation of intellectual property rights.[77]

ENFORCEMENT OPTIONS - The Twitter Rules (along with all incorporated policies), Privacy Policy, and Terms of Service (TOS) collectively make up the "Twitter User Agreement" that governs a user's access to and use of Twitter's services.

Failure to comply with Twitter's rules may result in one or more enforcement actions, such as:

- Temporarily limiting the user's ability to create posts or interact with other Twitter users;
- Requiring the user to remove prohibited content before they can again create new posts and interact with other Twitter users;
- Asking the user to verify account ownership with a phone number or email address; or
- Permanently suspending the user's account(s).[78]

[67] Colin Crowell and Mahima Kaul, Protecting the integrity of the election conversation in India, 21 Feb. 2019, Twitter, https://blog.twitter.com/en_in/topics/events/2019/election-integrity.html

[68] *Ibid*

[69] Hateful Conduct Policy, Twitter, https://help.Twitter.com/en/rules-and-policies/hateful-conduct-policy

[70] Updating our Rules against Hateful Conduct, Twitter, https://blog.Twitter.com/official/en_us/topics/company/2019/hatefulconductupdate.html

[71] *Ibid*

[72] Abusive Behaviour, Twitter, https://help.Twitter.com/en/rules-and-policies/abusive-behavior?lang=browser

[73] Violent Threats Policy, Twitter, https://help.Twitter.com/en/rules-and-policies/violent-threats-glorification

[74] Glorification of Violence Policy, Twitter, https://help.Twitter.com/en/rules-and-policies/glorification-of-violence

[75] Private Information Policy, Twitter, https://help.Twitter.com/en/rules-and-policies/personal-information

[76] Sensitive Media Policy, Twitter, https://help.twitter.com/en/rules-and-policies/media-policy

[77] The Twitter Rules, Twitter, https://help.twitter.com/en/rules-and-policies/twitter-rules

[78] Twitter Rules Enforcement, Twitter, https://transparency.twitter.com/en/twitter-rules-enforcement.html

Troll Patrol India: Recommendations

Online abuse against women on this scale does not have to exist on Twitter. The company's failure to adequately meet its human rights responsibilities regarding online abuse will continue to silence women on the platform unless Twitter undertakes, with urgency, concrete steps to effectively tackle this problem.

SUMMARY OF RECOMMENDATIONS:

- Twitter should publicly share comprehensive, meaningful and disaggregated information about the nature and levels of online abuse against women on a country by country basis, as well as other groups, on the platform, and how they respond to it.
- Twitter should improve its reporting mechanisms to ensure consistent application and better response to complaints of violence and abuse.
- Twitter should provide more clarity about how it interprets and identifies violence and abuse on the platform and how it handles reports of such abuse.

RECOMMENDATIONS:

AMNESTY INTERNATIONAL INDIA IS ASKING TWITTER TO:

1. Publish meaningful data on how they handle online abuse, including:

a) Disaggregated data of nature of abuse and hateful conduct reported by users, disaggregated by country, including whether the hateful conduct was based on race, ethnicity, caste-status, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability or serious disease.

b) Action taken on the aforementioned reports of abuse published on an annual basis, with country–specific break up.

c) The criteria employed to adjudicate violations and decide penalties, and the time taken for the adjudication process.

d) Information regarding the criteria and decision for granting appeals (or not), year and countrywise number of appeals received, with outcomes.

e) Clarify that the standards of scrutiny are same for verified and non-verified account users in cases of reports of abuse against them and publish information on the same.

2. Improve reporting mechanisms by ensuring:

a) Feedback is given to the person reporting abuse, on the action taken and whether the user is satisfied with the action taken by Twitter. The data gathered through this should be published.

b) Published data on reports of abuse, their outcomes and user satisfaction with the processes is used to ensure that there is no discrimination on part of content moderators, based on gender, caste, religion, ethnicity, race, marital status, age and other such identifying criteria.

c) Effectiveness of the reporting mechanisms is evaluated by measuring whether abuse is in effect reducing or not.

3. Provide more clarity about Twitter's internal mechanisms related to content moderation and language by:

a) Publishing further details on Twitter's language detection mechanism and how it detects abuse in regional languages, colloquial slang, mixed language tweets (Hindi and English mixed), or languages where native scripts are used alongside Latin scripts.

b) Sharing and publishing the number of content moderators Twitter employs, including the number of moderators employed per region and by language along with information on the volume of content handled and time allocated, per moderator.

c) Sharing how moderators are trained to identify gender and other identity-based violence against users, about international human rights standards, Twitter's responsibility to respect the rights of users on its platform, including the right for women to express themselves on Twitter freely and without fear of violence and abuse and language related trainings.

d) Providing details about any automated processes used to identify online abuse against women, detailing technologies used, accuracy levels, any biases identified in the results and information about how (if) the algorithms are currently on the platform.

e) Sharing information on how language is detected and abusive tweets are identified in images that contain abusive text and are circulated on the platform.

4. Improve security and privacy features by:

a) Enhancing the existing safety and awareness campaigns amongst all users about the harmful impact of online abuse on the platform.

Annexure 1: Enrichment

In addition to the sample of tweets selected at random from the Twitter data obtained from our third-party tool, we included a smaller sample of enriched tweets in the data shown to Decoders. That is, a sample of tweets that have been selected to potentially contain a larger proportion of abusive content. This was done only to enhance the experience of Decoders, by giving them a higher probability to identify abuse and not for the final analysis and estimates presented in this study.

BUILDING THE ENRICHMENT MODEL

We decided to build a Naive Bayes model to identify tweets that may be abusive. We had multiple types of labelled input data which included:

- ~1500 labelled tweets by Amnesty International India staff: This is the only input data that reflects the real data and therefore the entire test set was pulled from here. The tweets were split to test set (30%) and training set component (70%) which was added to the below sources, classed 100% as training.
- A set of Hindi swear words and offensive phrases, some of which had English translations, were sourced from the labelled set. All, barring one, were Latin spelling, which we transliterated to Devanagari using the R stringi package, so that tweets in Devanagari would also be labelled. Some software transliterations were found to be incorrect. All these abuse terms were treated as individual samples (tweets) with offensive label = 1. An alternative would have been using them instead as keywords within real tweets. A test on a small 200-tweet sample, showed that this approach did not increase abusive tweet proportion, so we used the "input sample" option instead.
- Indian vocabulary was sourced from Hatebase, a hate speech detection platform. At the time of extraction this consisted of about 30 terms across all non-English languages relevant to the project. It included Latin and native scripts. All of these were treated as individual samples (tweets) with offensive label = 1. We did not use English terms as they included common words that can be offensive in certain contexts, largely in the USA.
- The Kaggle "Detecting Insults in Social Commentary" data (test and training sets combined). This is a freely available online dataset of comments labelled for offensiveness. They are mostly in English. They were also used in the original Troll Patrol enrichment model training.While not an exact match for the 'problematic' or 'abusive' tweets that we are targeting, considering the different culture, time, etc., this option was thought to be valuable.
- We also tried to use some labelled English tweets from the previous Troll Patrol project. When combined with the rest of this data, it gave very poor predictivity in our sample (eg. 90% positive for abuse). We had limited time to explore this possibility and did not know if the poor predictivity was because we had equal numbers of abusive and non-abusive tweets in the training set and it biased the data. Due to time constraints on exploring, we did not use this data.

BUILDING AND TRAINING THE NAIVE BAYES MODEL

The enrichment model used the presence of different terms as signifying abuse.

Language processing steps - We made all the text into lower case, removed punctuation and numbers, stripped extra spaces, and removed English stop words (Hindi stop words was not available). We experimented with stemming words in English, but it did not help performance and so it was not done.

After preparing the corpus and document term matrix (DTM), we separated the labelled set into test and training parts. The test data comprised 30% of the labelled tweets from the Amnesty International India team. A full model build would have probably included a validation set including these to check for underfitting.

We however, did not run this check, because the sole purpose of the Naive Bayes model was to find more abusive samples than average for the use by Decoders' and not for actual content moderation or prediction.

We then limited the word dictionary to those with a certain number of appearances in the training data (we tried 3-6, and used 5 appearances as the threshold). As such, we did not use vocabulary with very low number of appearances in training data, since it was unlikely to be strongly predictive in the test set, and including low frequency words would have increased the model training time.

The next step was to set up a matrix showing the presence or absence of each word in each tweet/sample. For this study, we used single-words only. Given more time and crucial performance requirements we would have tried bigrams and trigrams (two-word and three-word sequences), in addition to investigating different models.

The classifier for offensiveness was trained on the training data.

MODEL PERFORMANCE ON TEST SET

In our sample, the performance of the model was reasonably poor in both sensitivity and specificity.

It increased abuse by only 1-5%, performing slightly better on non-English than English tweets in the test set.

Group	Improvement	Base positive rate (test set)
Overall	3.8%	13.1%
English	1.5%	7.7%
Non-English	5.1%	16.5%

Despite the poor performance, the model was used to enrich the sample of tweets seen by Decoders. This was done to enhance, even if only marginally, the user experience by increasing the likelihood of them encountering problematic or abusive tweets. The poor performance of the model meant that enrichment worked only in minority languages (non-English and not Hindi). This was likely because of the Hatebase sample and the presence of very few minor language tweets in the sample.

Annexure 2: Languages, Language Detection and Flagging

This study included tweets in 8 Indian languages other than English.

For practical reasons we could not obtain Twitter's own language classification via Crimson Hexagon (the tool used to obtain the Tweet sample analysed by Decoders). The tweets we obtained from Crimson Hexagon were therefore not weighted by language and we opted to detect language using other third party services in the second stage.

DETECTING LANGUAGES

The simplest tool available to us for language detection was the Google language detection service embedded in Google Sheets (`=detectlanguage()`). This was only applied to our Decoders tweet samples rather than the full random sets. To determine the language categories to use in our crowd-sourced project, we investigated our sample of tweets to be provided in the first Decoders set, and found 8 Indian languages and English with 1% or more of the tweet volume.

In decreasing order of occurrence in this sample, they were: Hindi, English, Marathi, Telugu, Tamil, Gujarati, Kannada, Bengali, and Malayalam. These languages were presented to Amnesty Decoders as options for selection.

This meant that some languages spoken in India with an even smaller Twitter presence could not be chosen by users – for instance, Urdu and Punjabi.

Rather than delete tweets that were not found to be in one of the above languages, we included them into the major language, Hindi. This meant that users selecting Hindi as an option were occasionally seeing a non-Hindi tweet, but it also meant that we had a comprehensive sampling of tweets rather than filtering out some because of what was known to be an imperfect detection tool.

Decoders could either flag these tweets as not being in a language they understood, which meant other Decoders on our forum could see the tweet, or – if they happened to be familiar with the Tweet language anyway – the user could identify problematic or abusive content themselves.

ISSUES IN LANGUAGE DETECTION

Some of the issues that came up with language detection:

- We only had access to the content of the tweet for language detection. We assume that Twitter has more sophisticated language detection tools, using for example the user sign-up language or location. Therefore, we believe that our language characterisation is likely less accurate than that of Twitter itself.
- Tweets are short and when only one or two words are available exclusive of mentions and hashtags, no language detection will be perfect.
- It was noted that some tweets were in “Hinglish” – a mix of Hindi and English – or in mixtures of other languages. This could especially be the case for items like slogans or news references. This made detection very difficult.
- Non-English languages could be used in Tweets in Latin script or native script (e.g. Devanagari for Hindi).
- Sometimes a tweet included images or videos in a different language to the text written, or with minimal text written. This meant that 'for example' users selecting English language tweets could see videos in Tamil. We know that our language detection was only focused on the text; it is unclear whether Twitter analyses text inside images.

ALTERNATIVE LANGUAGE DETECTION

R PACKAGE ‘FRANC’

‘franc’ is a free R package for language detection. We investigated its use but testing revealed it was not especially accurate. The designers specified that it works better on longer text, and without access to the latest models via API it was likely to suit poorly to Twitter language detection that uses slang and new words.

MICROSOFT TRANSLATE API

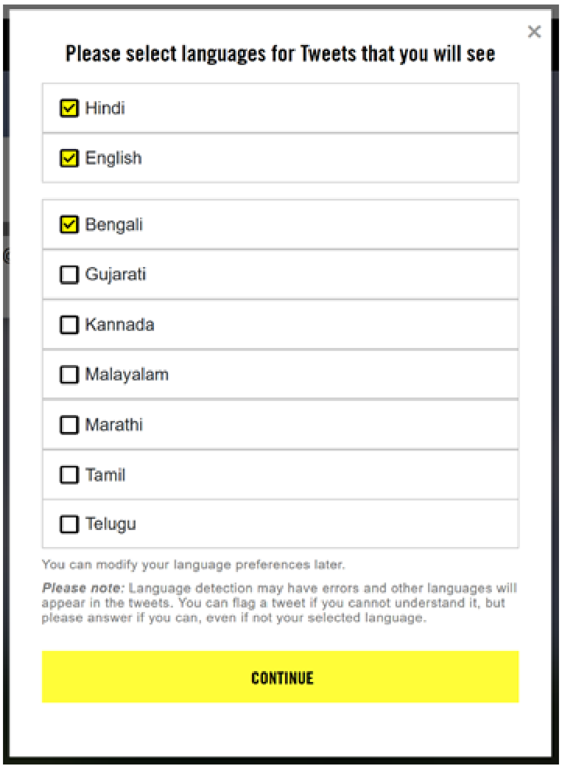
We investigated using the Microsoft Translate API, which includes detection options. The API does have a cost based on characters detected/ translated, so this was used on a data subset under the free plan first.

We did not pursue this option because we did not see great model performance improvements. A small-sample testing indicated that the Microsoft system seemed less good at identifying languages in different script (e.g. Hindi written in Latin characters) than the Google's Detectlanguage() function.

LANGUAGE PRESENTATION IN DECODERS PLATFORM

Decoders were given language options before they started decoding tweets – they could select more than one language and change their language choices at any time.

Language selection options on the Amnesty Decoders platform



Please select languages for Tweets that you will see

☑ Hindi
☑ English

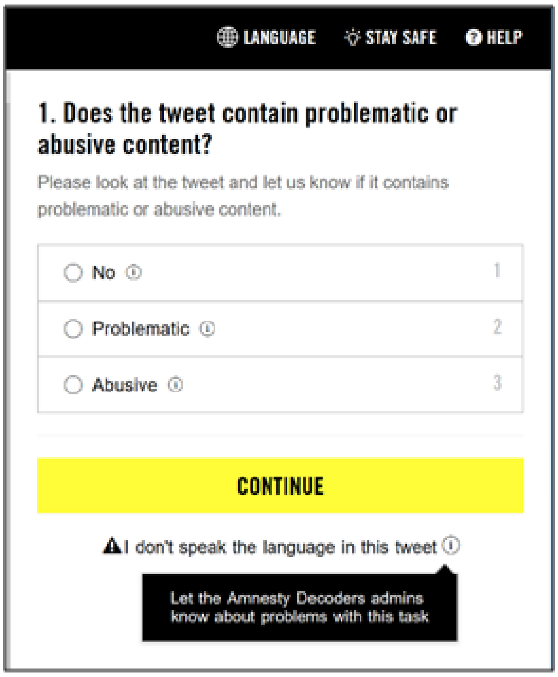
☑ Bengali
☐ Gujarati
☐ Kannada
☐ Malayalam
☐ Marathi
☐ Tamil
☐ Telugu

You can modify your language preferences later.

Please note: Language detection may have errors and other languages will appear in the tweets. You can flag a tweet if you cannot understand it, but please answer if you can, even if not your selected language.

If the user did not understand the language of a tweet presented, they could flag that specific task. This means that the tweet was sent to the discussion forum, with an auto-generated note mentioning the language identified for the tweet.

Option for users to flag tasks where the language of the tweet was potentially missclassified



1. Does the tweet contain problematic or abusive content?

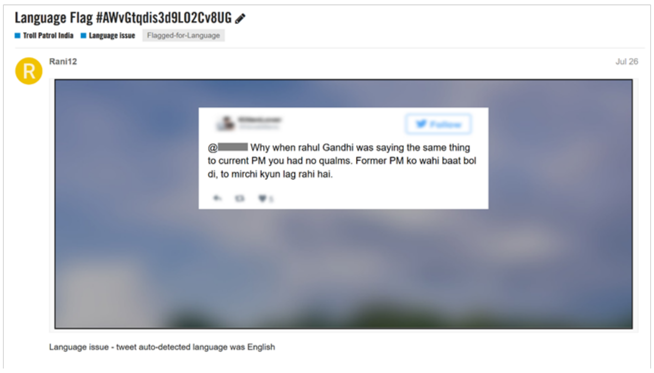
Please look at the tweet and let us know if it contains problematic or abusive content.

○ No 1
○ Problematic 2
○ Abusive 3

I don't speak the language in this tweet

Let the Amnesty Decoders admins know about problems with this task

Example of a tweet flagged for language misclassification shown in the discussion forum.



Language Flag #AWvGtqdis3d9L02Cv8UG

Rani12 Jul 26

@[redacted] Why when rahul Gandhi was saying the same thing to current PM you had no qualms. Former PM ko wahi baat bol di, to mirchi kyun lag rahi hai.

Language issue - tweet auto-detected language was English

Community Manager Volunteers, who were also Decoders in addition to managing the discussion forum, gave feedback and educated users on the forum on flagging for language. For instance, they re-emphasised that the goal was to determine abusiveness of tweets and not language. This meant that if Decoders could understand the tweet, which was in a language other than that selected by her/him, they were encouraged to nevertheless answer the question.

LANGUAGE FEEDBACK FLAGGING

Of the 142,474 tweets with any decoder answers, 4,370 had at least one decoder flag them for language issues.

Some languages had many more flags than others. Malayalam had 21.2% of tweets flagged and Kannada had 17.9%, while Tamil was also high with 11.2%. In contrast, English and Hindi had only 2% of tweets flagged.

This suggests that the language detection tool is better for some languages than others, and also that users were more familiar with some languages than others.

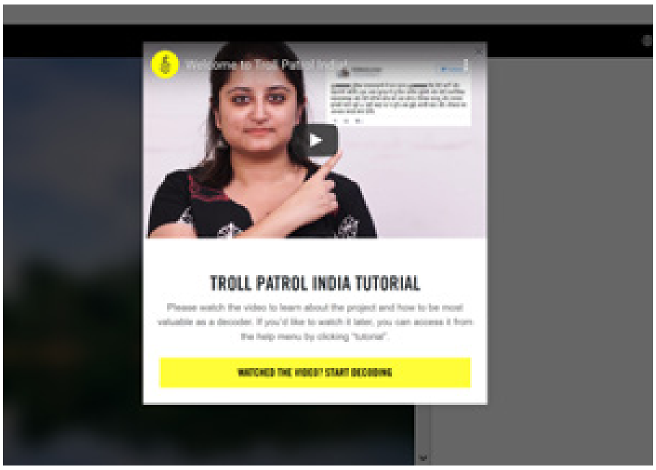
Proportion of tweets flagged for language missclassification for each detected language

Detected language	Proportion flagged
Bengali	6%
English	2%
Gujarati	7.1%
Hindi	2%
Kannada	17.9%
Malayalam	21.2%
Marathi	5.8%
Tamil	11.2%
Telugu	3.7%

As expected, there was more flagging for language when tweets contained videos or images (7.74%) than text alone (2.2%). This would be because images and videos also contain a different language than the tweet body.

Annexure 3: Amnesty Decoders Tool and Screenshots

1 Welcome screen and video tutorial

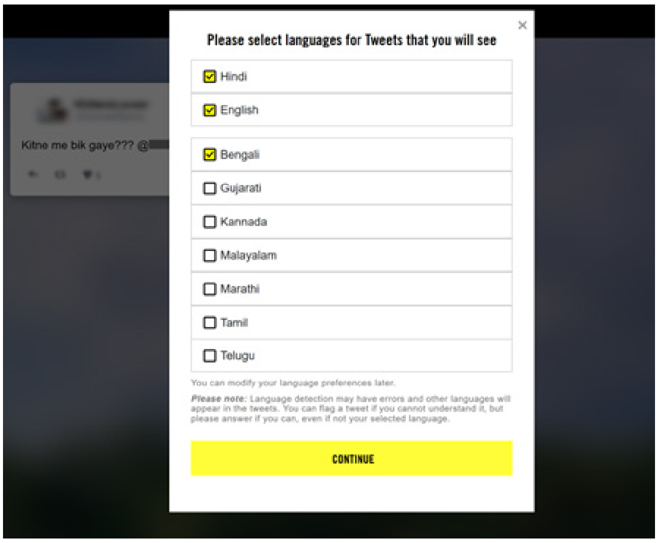


TROLL PATROL INDIA TUTORIAL

Please watch the video to learn about the project and how to be most valuable as a decoder. If you'd like to watch it later, you can access it from the help menu by clicking "tutorial".

WATCHED THE VIDEO? START DECODING

2 Language selection



Kitne me bik gaye??? @[redacted]

Please select languages for Tweets that you will see

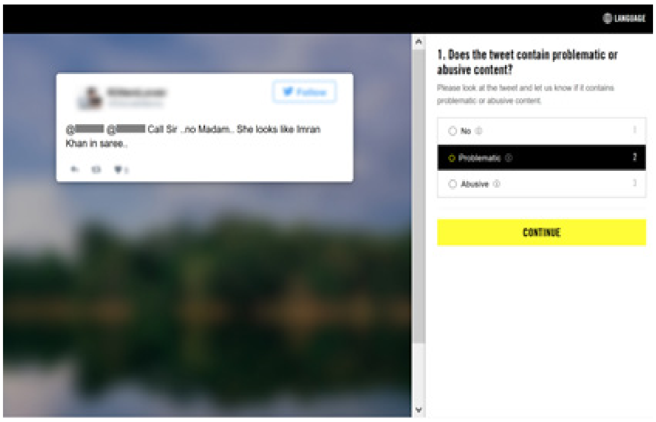
☑ Hindi
☑ English
☑ Bengali
☐ Gujarati
☐ Kannada
☐ Malayalam
☐ Marathi
☐ Tamil
☐ Telugu

You can modify your language preferences later.

Please note: Language detection may have errors and other languages will appear in the tweets. You can flag a tweet if you cannot understand it, but please answer if you can, even if not your selected language.

CONTINUE

3 Question one: does the tweet contain problematic or abusive content?



@[redacted] @[redacted] Call Sir ..no Madam.. She looks like Imran Khan in saree..

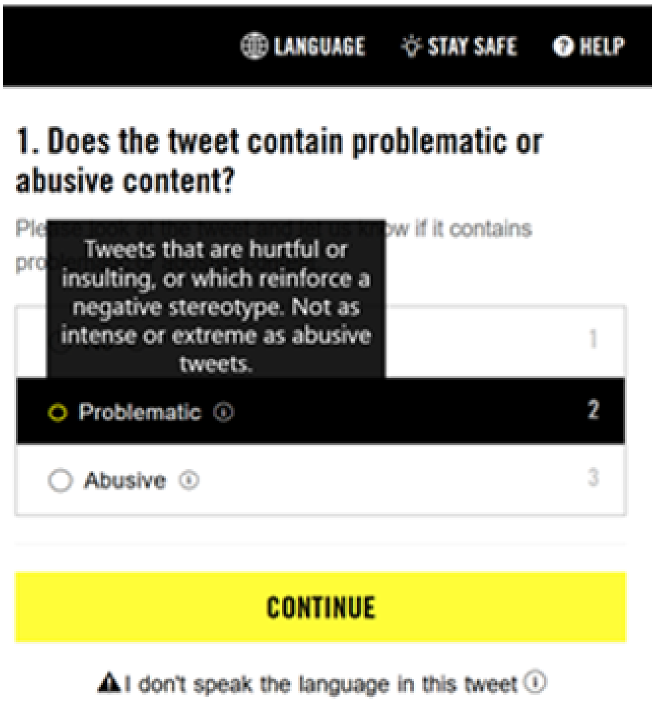
1. Does the tweet contain problematic or abusive content?

Please look at the tweet and let us know if it contains problematic or abusive content.

○ No 1
● Problematic 2
○ Abusive 3

CONTINUE

4 Short definitions available via tooltip in the question itself



1. Does the tweet contain problematic or abusive content?

Ple w if it contains pro

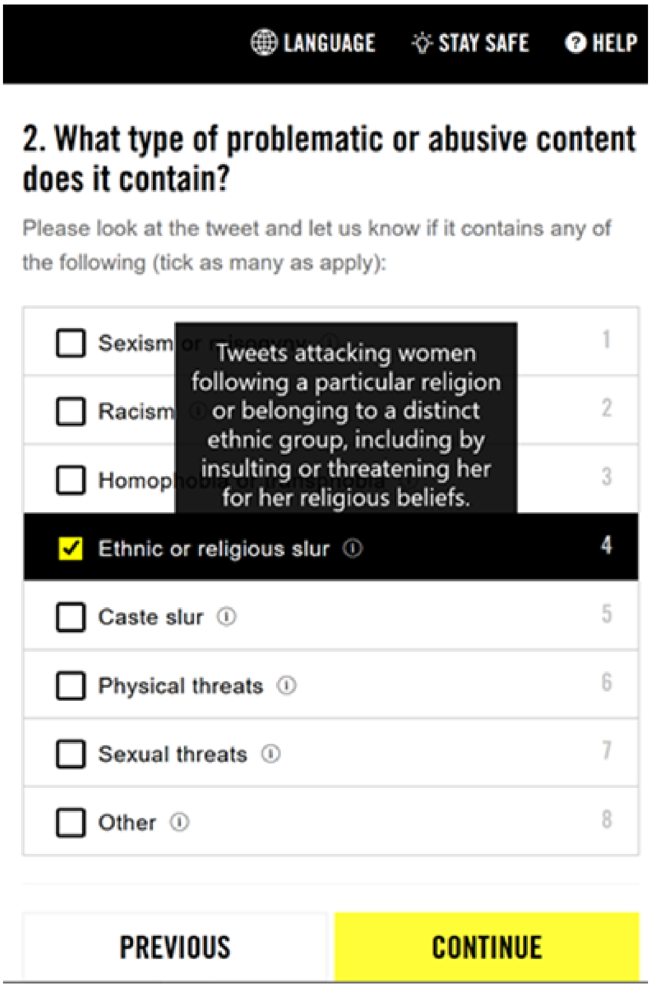
Tweets that are hurtful or insulting, or which reinforce a negative stereotype. Not as intense or extreme as abusive tweets.

1
● Problematic 2
○ Abusive 3

CONTINUE

⚠ I don't speak the language in this tweet

5 Question two: type of abusive or problematic content and tooltip support



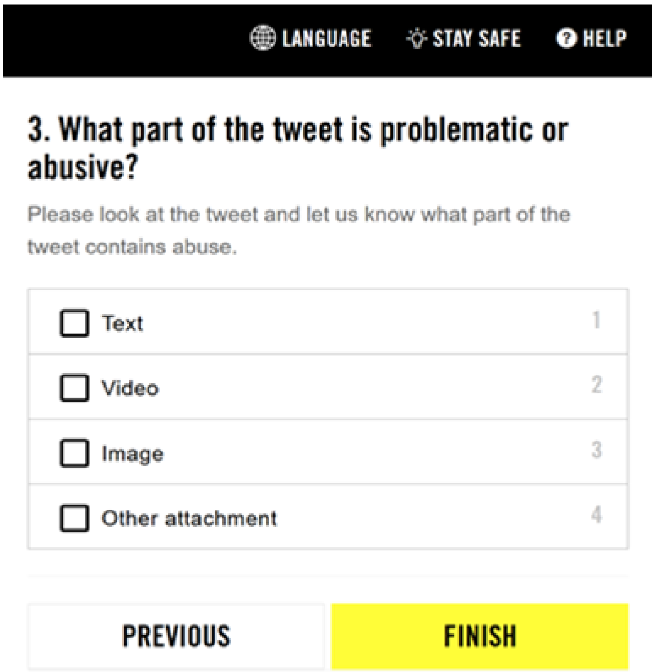
2. What type of problematic or abusive content does it contain?

Please look at the tweet and let us know if it contains any of the following (tick as many as apply):

- □ Sexism or misogyny ⓘ 1
- □ Racism ⓘ 2
- □ Homophobia or transphobia 3
- ☑ Ethnic or religious slur ⓘ 4
- □ Caste slur ⓘ 5
- □ Physical threats ⓘ 6
- □ Sexual threats ⓘ 7
- □ Other ⓘ 8

Tweets attacking women following a particular religion or belonging to a distinct ethnic group, including by insulting or threatening her for her religious beliefs.

6 Question 3: medium of abuse

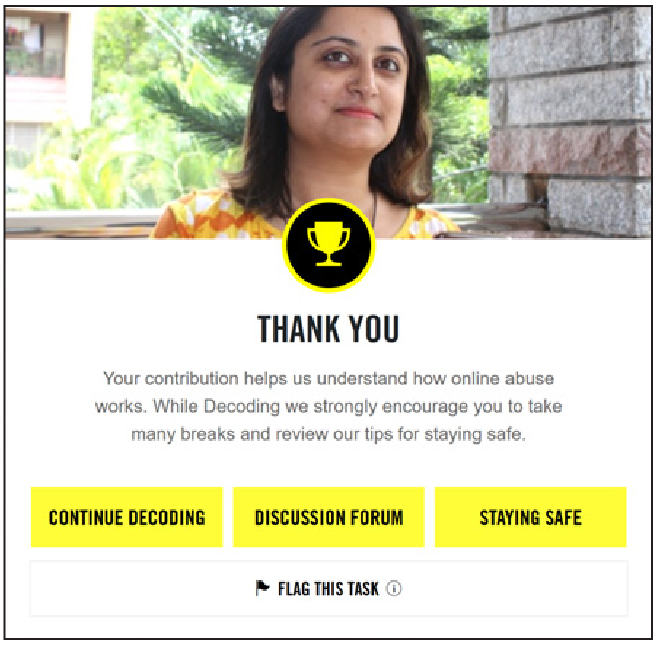


3. What part of the tweet is problematic or abusive?

Please look at the tweet and let us know what part of the tweet contains abuse.

- □ Text 1
- □ Video 2
- □ Image 3
- □ Other attachment 4

7 Thank you screen



THANK YOU

Your contribution helps us understand how online abuse works. While Decoding we strongly encourage you to take many breaks and review our tips for staying safe.

Annexure 4: Weighting

In order to analyse assignment answers, we re-weighted answers to account for sample mismatch with the pool of tweets (the true distribution in the "world set"). Weighting factors used applied to both responses and tweets:

RESPONSE WEIGHTING

While the target was three responses per tweet, some tweets had 10s of responses, some had only one or two. So as to represent the true distribution of abuse we needed to weight these answers. (This meant that one response of 'Abusive' out of three annotations for a tweet, was equivalent to two responses of 'Abusive' out of six annotations: both were equally abusive).

TWEET WEIGHTING

Day weighting

We selected tweets in proportion to the total number of tweets to these women per day. That is, a day in which 100,000 tweets were received would have about twice as many tweets in our sample as a day that had 50,000 tweets mentioning these women. Given that the project launched mid-way through the period of tweet collection (in order to take advantage of focused interest in the Indian elections) we did not have final day volumes for the day weighting at the time the project started. We, therefore reweighted the tweets after the project to reflect the true day weightings.

More precisely, the tweets were obtained in two "batches" (pre-election period [batch 1] and during and post-election [batch 2]) each with different proportions of tweets per day. We, therefore reweighted per batch - knowing that day weights within each batch were proportionate to the total tweets received by the women.

Language distribution weighting

In the final batch of tweets, we did not have decoding completed – and the language proportions were not equivalent. For instance, all English tweets were completed but not all Hindi. We weighted each tweet by language to match the original (as-detected) random set distribution of languages, so as not to skew results.

TWEET WEIGHTING STEPS

1. We found the "world set" distribution of batches from Daily Volumes data from Crimson Hexagon, p_W (b). This showed the total number of tweets per day that mentioned the women politicians of interest. Since we had already sampled in proportion to daily tweets within each batch, we only needed to reweight by batch, not by day.

2. We found the "world set" language probability distribution from random tweets per batch. For the second batch, we needed to use random set as uploaded (R') which included our final upload set of tweets on the Decoders platform. This set was randomly sampled but not completed, and more English tweets were decoded than Hindi. This meant that the tweets as completed did not show the true language distribution.

3. By applying the language probabilities and the batch probabilities, we found the probability for each batch-language combination within the "world set" of tweets. This gave us p_R' (l,b).

4. We applied the target probabilities to each language-batch combination as sampled, to reweight the entire Decoders random tweet set A.

$$w(l,b) = \frac{p_{R'}(l,b)}{p_A(l,b)}$$

This allowed us to use all random tweets.

After taking language distributions within batches, we used the daily volumes from Crimson Hexagon to set the batch-weighting. For our two time periods (1 March 2019 - 10 April 2019 and 11 April 2019 - 31 May 2019) the proportions in the "world set" were 44% to 56%. In our decoding sample because we did not have weighting factors in advance, it was 24% to 76%. This meant that tweets from the first-time period needed to be weighted more heavily so as to make sure that the time period was not underrepresented in our results.

Annexure 5: Agreement Analysis

Agreement amongst raters of tweets were quantified using two measures: Fleiss’ kappa (Fleiss, 1971) and intra-class correlation coefficient (Shrout, 1979).

FLEISS’ KAPPA

Fleiss’ kappa is a statistical measure of the degree of agreement among multiple raters classifying items (e.g. if a tweet contains problematic or abusive content: 'No', 'Problematic', or 'Abusive').

The measure calculates the degree of agreement above chance level. Values of the kappa statistic may cautiously be interpreted as follows (Landis and Koch, 1977), but it should be noted that the value is also affected by the number of classes and of items classified (Sim and Wright, 2005).

Kappa	Agreement
<0	Poor
0.01-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

A shortcoming of Fleiss’ kappa is that it treats the classes as nominal (i.e. non-ranked or non-ordinal) and hence disagreement is the same between all class pairs independent of natural class ranking. For example, disagreement between 'No' and 'Abusive' is considered the same as disagreement between 'Problematic' and 'Abusive' although disagreement is larger for the former. When used with ranked or ordinal classes, Fleiss’ kappa therefore tends to underestimate agreement.

INTRA-CLASS CORRELATION COEFFICIENT

The intra-class correlation coefficient (ICC) is another statistical measure of the degree of agreement between raters and it accounts for ordinal nature of classes, for example, the ranking 'No' 'Problematic', 'Abusive' by intensity. Intuitively, the measure evaluates agreement in terms of the proportion of the overall variation in classifications that is explained by inter-item variation (e.g. between tweets: one tweet classified as 'No' and another tweet classified as 'Abusive') as opposed to intra-item variation (e.g. within tweets: one rater classifying a tweet as 'No' and another rater classifying the same as 'Abusive'). ICC values may be interpreted as follows (Cicchetti, 1994).

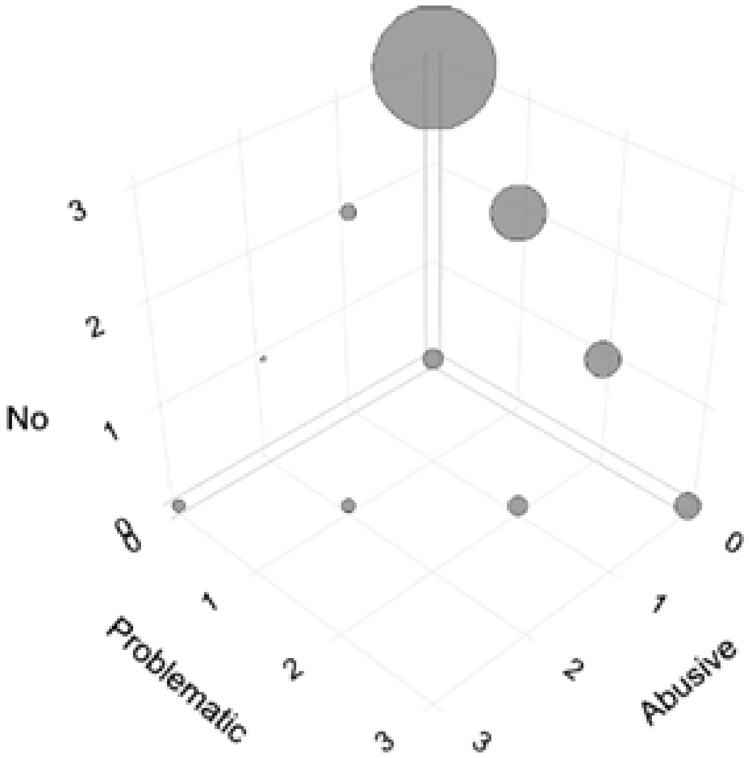
ICC	Agreement
<0.40	Poor
0.40-0.59	Fair
0.60-0.74	Good
0.75-1.00	Excellent

AGREEMENT MEASURES FOR EXPERT ANALYSIS

There was fair agreement of experts on whether tweets contained problematic, abusive or neutral content and good agreement when the ordinal nature of the classes was considered. The expert agreement was moderate, when 'Problematic' and 'Abusive' classes were grouped together. Expert agreement was fair, with respect to the type of 'Problematic' or 'Abusive' content.

Measure	Classification	Filter	Value
Kappa	Nominal: "No", "Problematic", "Abusive"	Tweets classified by at least 2 experts	0.37
Kappa	Nominal: "No", "Problematic or Abusive"	Tweets classified by at least 2 experts	0.43
Kappa	Nominal: Type of abuse, e.g. "Racism"	Tweets classified as "Problematic" or "Abusive" by at least 2 experts	0.27
ICC	Ordinal: "No"=0, "Problematic"=1, "Abusive"=2	Tweets classified by 3 experts	0.74

Tweet distribution by number of experts per class



This visualisation shows the distribution of tweets with respect to the number of experts that classified the individual tweets as 'No', 'Problematic', or 'Abusive'. The underlying data of the visualisation is equivalent to that used in the calculation of the intra-class correlation coefficient - that is, each of the tweets has classifications from exactly three experts. Therefore, the sum of the values of the 'No', 'Problematic', and 'Abusive' axes is three for each point.

The points at the top, bottom-left, and bottom-right corners correspond to perfect agreement amongst the three expert raters. Here, all three experts classified the tweet as either 'No', 'Problematic', or 'Abusive', respectively. The point in the middle corresponds to complete disagreement between the three experts and the remaining points denote the tweets for which there was partial agreement. The size of the points identifies the proportion of tweets that had that particular combination of expert classifications. Hence, the higher the overall agreement, the larger the points at the top, bottom-left, and/or bottom-right corners.

There is perfect expert agreement for 74.5%, partial expert agreement for 23.7%, and no expert agreement for 1.8% of the tweets.

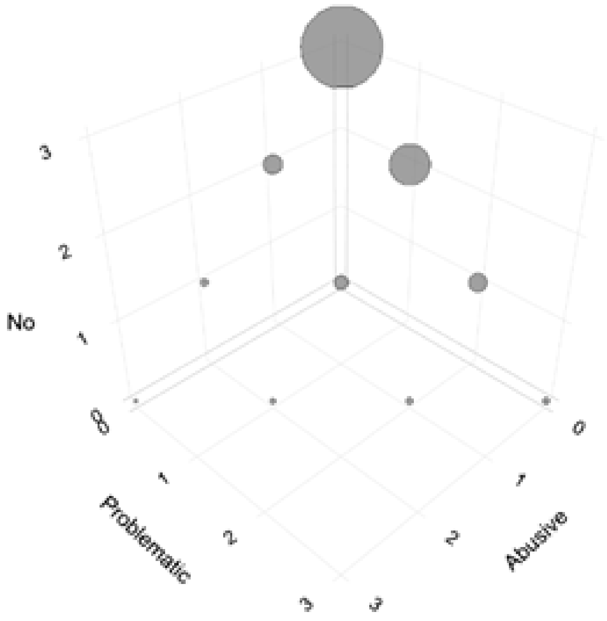
AGREEMENT MEASURES FOR DECODER ANALYSIS

As to be expected, agreement of Decoders was lower than that of experts. There was slight agreement amongst Decoders on whether tweets contain problematic, abusive or neutral content, but agreement was at least fair when the ordinal nature of the classification was taken into consideration. Decoder agreement was also fair when 'Problematic' and 'Abusive' classes were grouped together and also when evaluating the type of problematic or abusive content for tweets classified as such.

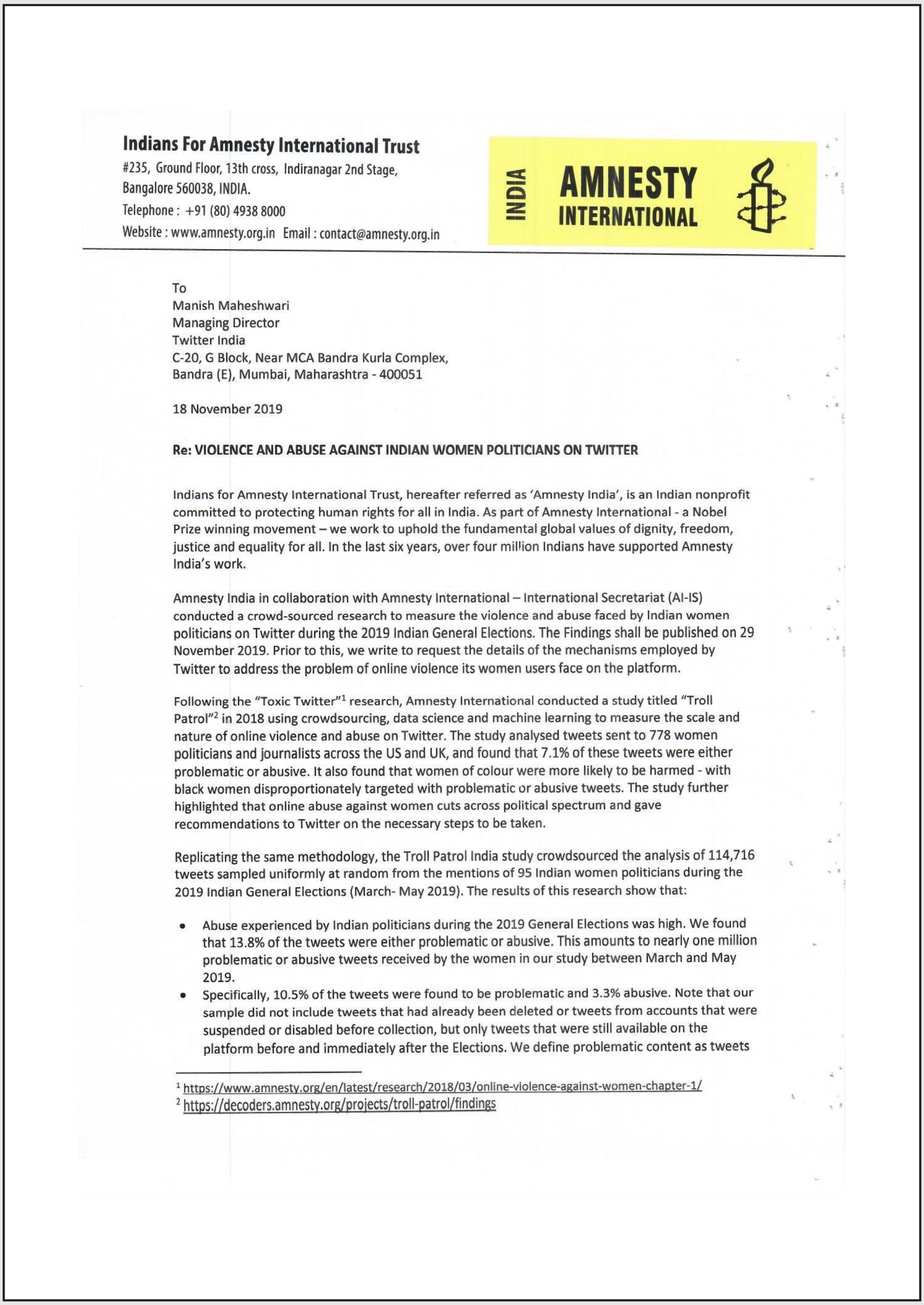
Measure	Classification	Filter	Value
Kappa	Nominal: “No”, “Problematic”,“Abusive”	Tweets classified by at least 2 Decoders	0.17
Kappa	Nominal: “No”, “Problematic or Abusive”	Tweets classified by at least 2 Decoders	0.20
Kappa	Nominal: Type of abuse, e.g. “Racism”	Tweets classified as “Problematic” or “Abusive” by at least 2 Decoders	0.22
ICC	Ordinal: “No”=0, “Problematic”=1, “Abusive”=2	Tweets classified by 3 Decoders	0.43

Tweet distribution by number of Decoders per class

There was perfect Decoder agreement for 71.2% of the tweets, partial decoder agreement for 26.6% of them, and no decoder agreement for 2.2%. In congruence with the findings regarding the intra-class correlation coefficients, the percentage of tweets with perfect agreement was lower for the Decoders than for the experts, while the percentage of tweets with partial or no agreement was higher.



Annexure 6: Letter to Twitter



Indians For Amnesty International Trust
#235, Ground Floor, 13th cross, Indiranagar 2nd Stage,
Bangalore 560038, INDIA.
Telephone : +91 (80) 4938 8000
Website : www.amnesty.org.in Email : contact@amnesty.org.in

INDIA AMNESTY INTERNATIONAL

To
Manish Maheshwari
Managing Director
Twitter India
C-20, G Block, Near MCA Bandra Kurla Complex,
Bandra (E), Mumbai, Maharashtra - 400051

18 November 2019

Re: VIOLENCE AND ABUSE AGAINST INDIAN WOMEN POLITICIANS ON TWITTER

Indians for Amnesty International Trust, hereafter referred as ‘Amnesty India’, is an Indian nonprofit committed to protecting human rights for all in India. As part of Amnesty International - a Nobel Prize winning movement – we work to uphold the fundamental global values of dignity, freedom, justice and equality for all. In the last six years, over four million Indians have supported Amnesty India’s work.

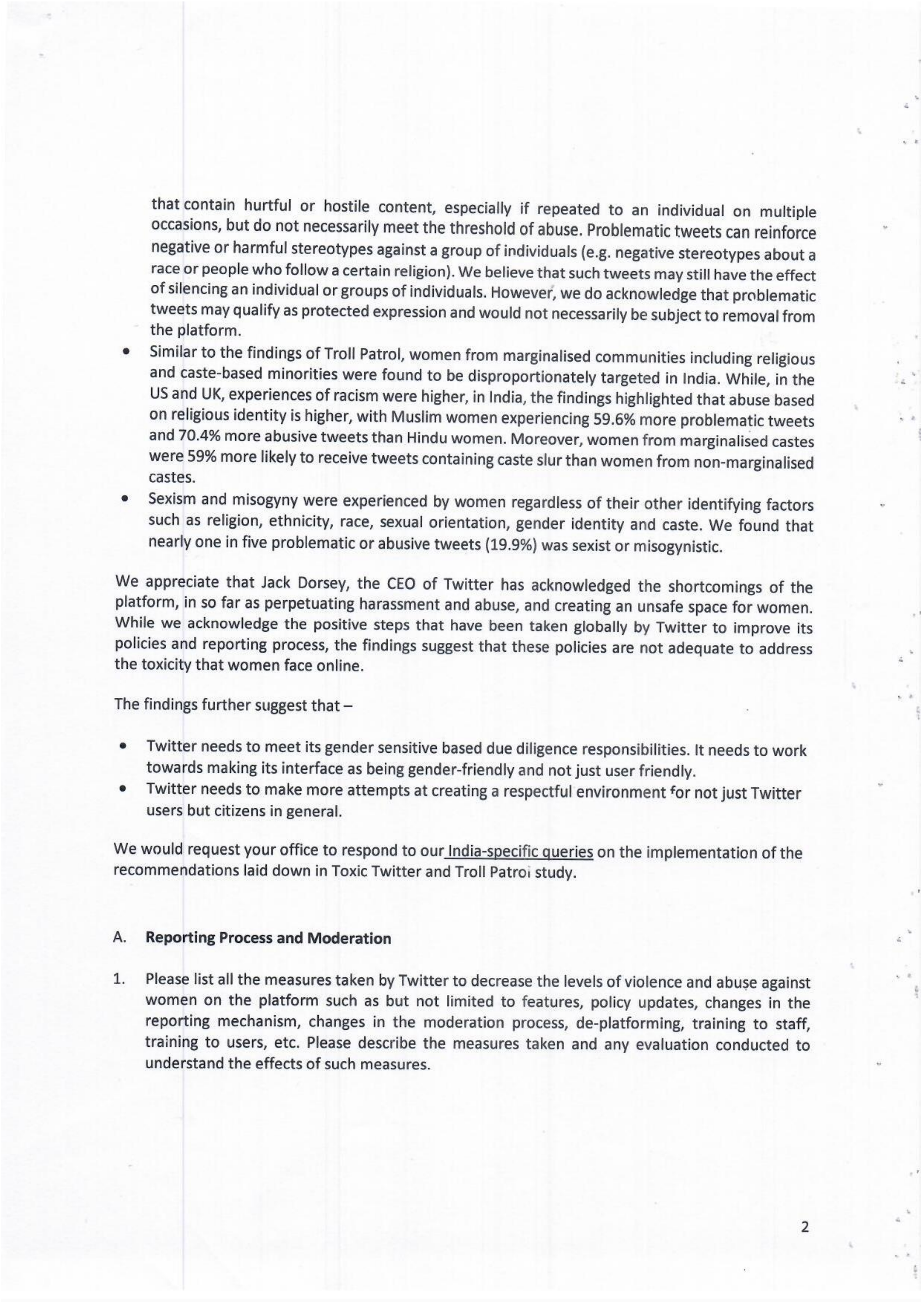
Amnesty India in collaboration with Amnesty International – International Secretariat (AI-IS) conducted a crowd-sourced research to measure the violence and abuse faced by Indian women politicians on Twitter during the 2019 Indian General Elections. The Findings shall be published on 29 November 2019. Prior to this, we write to request the details of the mechanisms employed by Twitter to address the problem of online violence its women users face on the platform.

Following the “Toxic Twitter”[1] research, Amnesty International conducted a study titled “Troll Patrol”[2] in 2018 using crowdsourcing, data science and machine learning to measure the scale and nature of online violence and abuse on Twitter. The study analysed tweets sent to 778 women politicians and journalists across the US and UK, and found that 7.1% of these tweets were either problematic or abusive. It also found that women of colour were more likely to be harmed - with black women disproportionately targeted with problematic or abusive tweets. The study further highlighted that online abuse against women cuts across political spectrum and gave recommendations to Twitter on the necessary steps to be taken.

Replicating the same methodology, the Troll Patrol India study crowdsourced the analysis of 114,716 tweets sampled uniformly at random from the mentions of 95 Indian women politicians during the 2019 Indian General Elections (March- May 2019). The results of this research show that:

- Abuse experienced by Indian politicians during the 2019 General Elections was high. We found that 13.8% of the tweets were either problematic or abusive. This amounts to nearly one million problematic or abusive tweets received by the women in our study between March and May 2019.
- Specifically, 10.5% of the tweets were found to be problematic and 3.3% abusive. Note that our sample did not include tweets that had already been deleted or tweets from accounts that were suspended or disabled before collection, but only tweets that were still available on the platform before and immediately after the Elections. We define problematic content as tweets

[1] https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1/
[2] https://decoders.amnesty.org/projects/troll-patrol/findings



that contain hurtful or hostile content, especially if repeated to an individual on multiple occasions, but do not necessarily meet the threshold of abuse. Problematic tweets can reinforce negative or harmful stereotypes against a group of individuals (e.g. negative stereotypes about a race or people who follow a certain religion). We believe that such tweets may still have the effect of silencing an individual or groups of individuals. However, we do acknowledge that problematic tweets may qualify as protected expression and would not necessarily be subject to removal from the platform.

- Similar to the findings of Troll Patrol, women from marginalised communities including religious and caste-based minorities were found to be disproportionately targeted in India. While, in the US and UK, experiences of racism were higher, in India, the findings highlighted that abuse based on religious identity is higher, with Muslim women experiencing 59.6% more problematic tweets and 70.4% more abusive tweets than Hindu women. Moreover, women from marginalised castes were 59% more likely to receive tweets containing caste slur than women from non-marginalised castes.
- Sexism and misogyny were experienced by women regardless of their other identifying factors such as religion, ethnicity, race, sexual orientation, gender identity and caste. We found that nearly one in five problematic or abusive tweets (19.9%) was sexist or misogynistic.

We appreciate that Jack Dorsey, the CEO of Twitter has acknowledged the shortcomings of the platform, in so far as perpetuating harassment and abuse, and creating an unsafe space for women. While we acknowledge the positive steps that have been taken globally by Twitter to improve its policies and reporting process, the findings suggest that these policies are not adequate to address the toxicity that women face online.

The findings further suggest that –

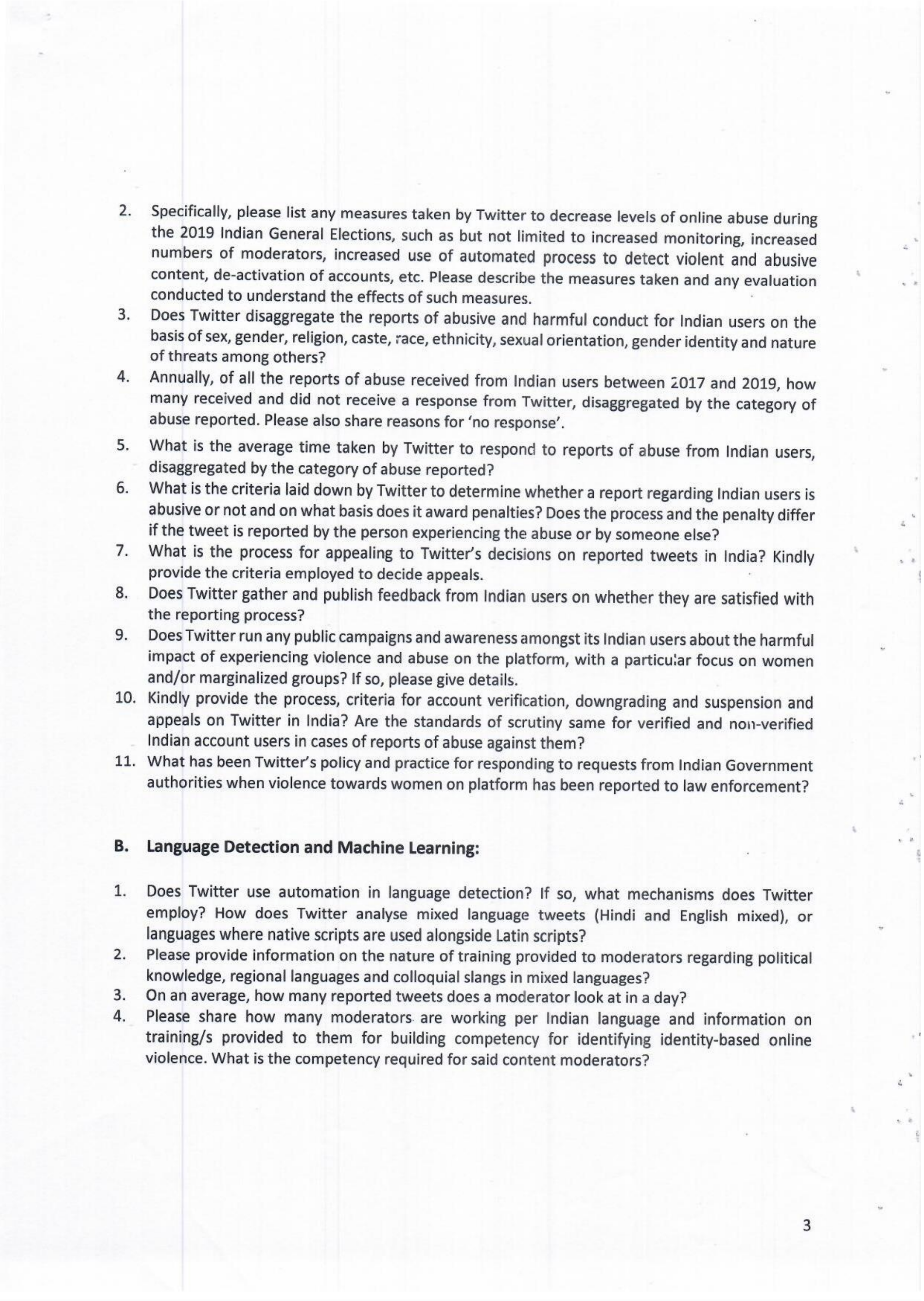
- Twitter needs to meet its gender sensitive based due diligence responsibilities. It needs to work towards making its interface as being gender-friendly and not just user friendly.
- Twitter needs to make more attempts at creating a respectful environment for not just Twitter users but citizens in general.

We would request your office to respond to our India-specific queries on the implementation of the recommendations laid down in Toxic Twitter and Troll Patrol study.

A. **Reporting Process and Moderation**

1. Please list all the measures taken by Twitter to decrease the levels of violence and abuse against women on the platform such as but not limited to features, policy updates, changes in the reporting mechanism, changes in the moderation process, de-platforming, training to staff, training to users, etc. Please describe the measures taken and any evaluation conducted to understand the effects of such measures.

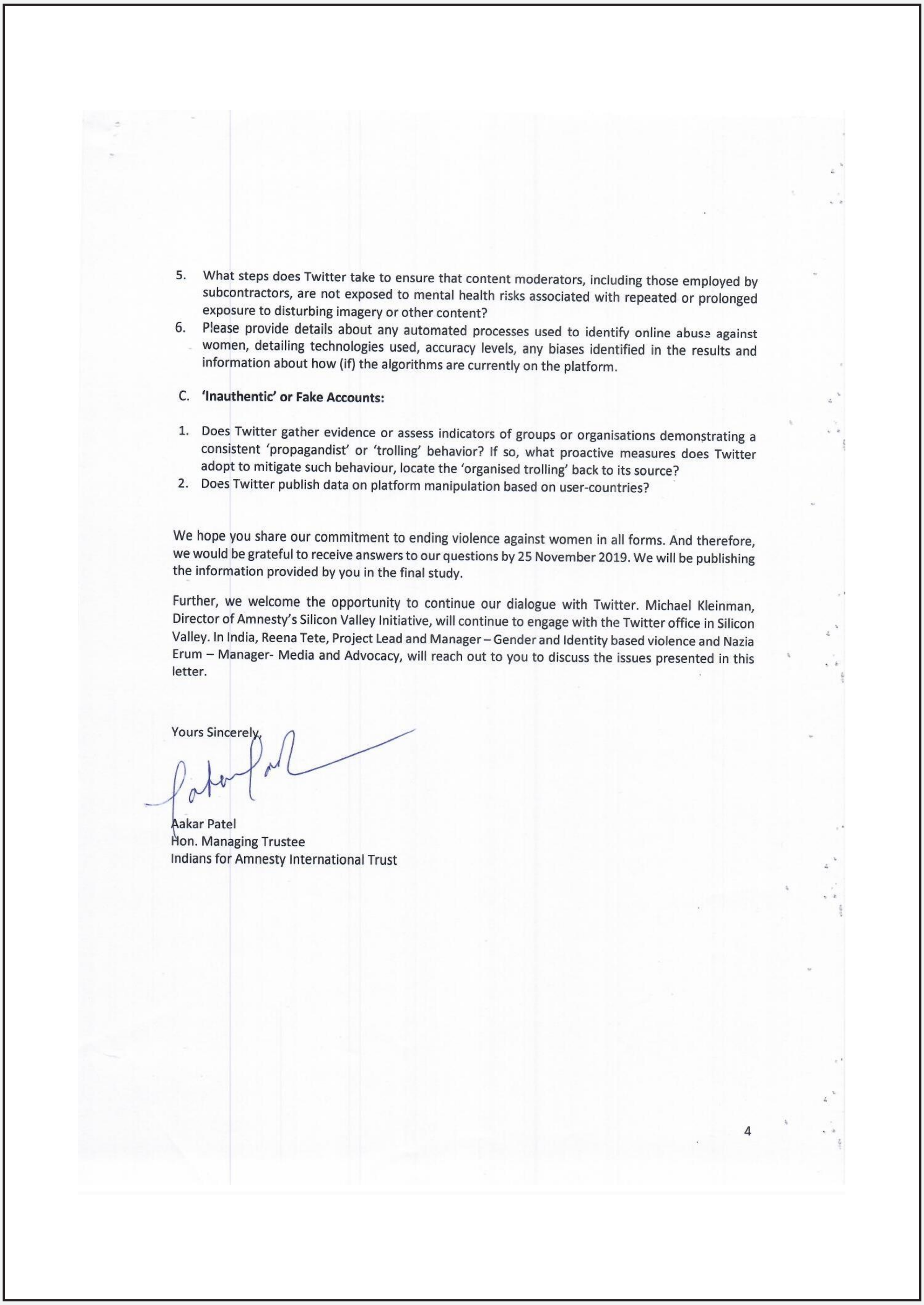
2



2. Specifically, please list any measures taken by Twitter to decrease levels of online abuse during the 2019 Indian General Elections, such as but not limited to increased monitoring, increased numbers of moderators, increased use of automated process to detect violent and abusive content, de-activation of accounts, etc. Please describe the measures taken and any evaluation conducted to understand the effects of such measures.
3. Does Twitter disaggregate the reports of abusive and harmful conduct for Indian users on the basis of sex, gender, religion, caste, race, ethnicity, sexual orientation, gender identity and nature of threats among others?
4. Annually, of all the reports of abuse received from Indian users between 2017 and 2019, how many received and did not receive a response from Twitter, disaggregated by the category of abuse reported. Please also share reasons for 'no response'.
5. What is the average time taken by Twitter to respond to reports of abuse from Indian users, disaggregated by the category of abuse reported?
6. What is the criteria laid down by Twitter to determine whether a report regarding Indian users is abusive or not and on what basis does it award penalties? Does the process and the penalty differ if the tweet is reported by the person experiencing the abuse or by someone else?
7. What is the process for appealing to Twitter's decisions on reported tweets in India? Kindly provide the criteria employed to decide appeals.
8. Does Twitter gather and publish feedback from Indian users on whether they are satisfied with the reporting process?
9. Does Twitter run any public campaigns and awareness amongst its Indian users about the harmful impact of experiencing violence and abuse on the platform, with a particular focus on women and/or marginalized groups? If so, please give details.
10. Kindly provide the process, criteria for account verification, downgrading and suspension and appeals on Twitter in India? Are the standards of scrutiny same for verified and non-verified Indian account users in cases of reports of abuse against them?
11. What has been Twitter's policy and practice for responding to requests from Indian Government authorities when violence towards women on platform has been reported to law enforcement?

B. Language Detection and Machine Learning:

1. Does Twitter use automation in language detection? If so, what mechanisms does Twitter employ? How does Twitter analyse mixed language tweets (Hindi and English mixed), or languages where native scripts are used alongside Latin scripts?
2. Please provide information on the nature of training provided to moderators regarding political knowledge, regional languages and colloquial slangs in mixed languages?
3. On an average, how many reported tweets does a moderator look at in a day?
4. Please share how many moderators are working per Indian language and information on training/s provided to them for building competency for identifying identity-based online violence. What is the competency required for said content moderators?



5. What steps does Twitter take to ensure that content moderators, including those employed by subcontractors, are not exposed to mental health risks associated with repeated or prolonged exposure to disturbing imagery or other content?
6. Please provide details about any automated processes used to identify online abuse against women, detailing technologies used, accuracy levels, any biases identified in the results and information about how (if) the algorithms are currently on the platform.

C. **'Inauthentic' or Fake Accounts:**

1. Does Twitter gather evidence or assess indicators of groups or organisations demonstrating a consistent 'propagandist' or 'trolling' behavior? If so, what proactive measures does Twitter adopt to mitigate such behaviour, locate the 'organised trolling' back to its source?
2. Does Twitter publish data on platform manipulation based on user-countries?

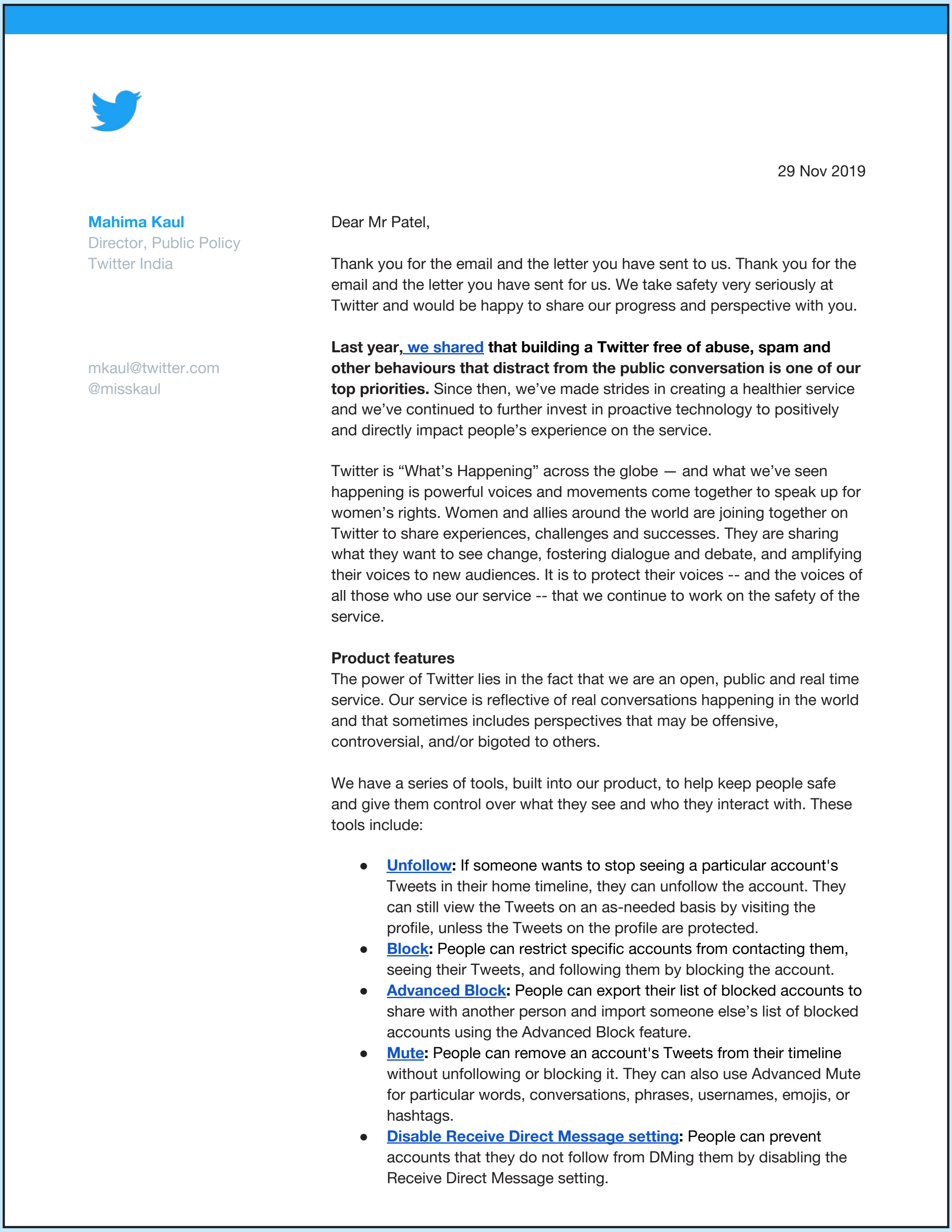
We hope you share our commitment to ending violence against women in all forms. And therefore, we would be grateful to receive answers to our questions by 25 November 2019. We will be publishing the information provided by you in the final study.

Further, we welcome the opportunity to continue our dialogue with Twitter. Michael Kleinman, Director of Amnesty's Silicon Valley Initiative, will continue to engage with the Twitter office in Silicon Valley. In India, Reena Tete, Project Lead and Manager – Gender and Identity based violence and Nazia Erum – Manager- Media and Advocacy, will reach out to you to discuss the issues presented in this letter.

Yours Sincerely,

Aakar Patel
Hon. Managing Trustee
Indians for Amnesty International Trust

Annexure 7: Twitter's Response



29 Nov 2019

Mahima Kaul
Director, Public Policy
Twitter India

mkaul@twitter.com
@misskaul

Dear Mr Patel,

Thank you for the email and the letter you have sent to us. Thank you for the email and the letter you have sent for us. We take safety very seriously at Twitter and would be happy to share our progress and perspective with you.

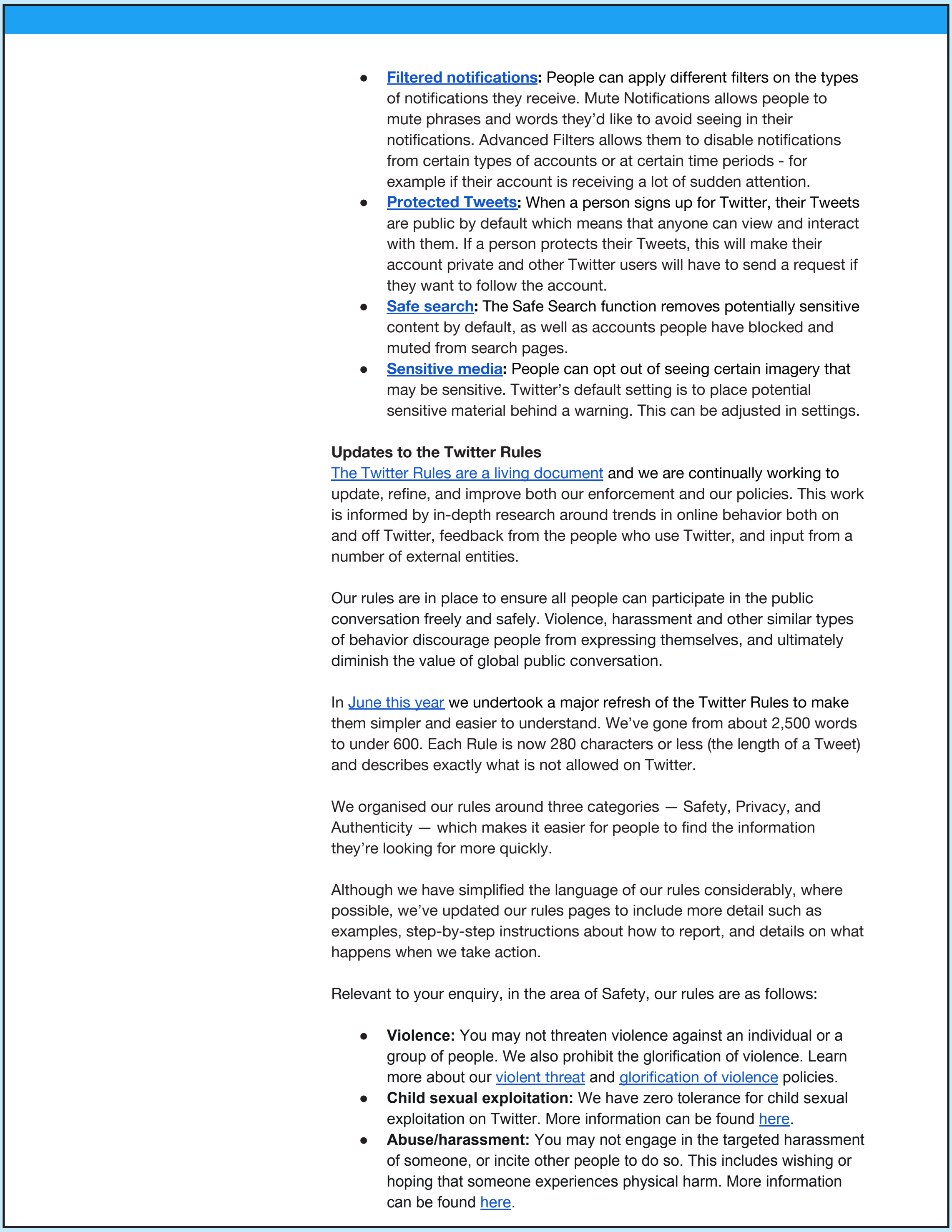
Last year, we shared that building a Twitter free of abuse, spam and other behaviours that distract from the public conversation is one of our top priorities. Since then, we've made strides in creating a healthier service and we've continued to further invest in proactive technology to positively and directly impact people's experience on the service.

Twitter is "What's Happening" across the globe — and what we've seen happening is powerful voices and movements come together to speak up for women's rights. Women and allies around the world are joining together on Twitter to share experiences, challenges and successes. They are sharing what they want to see change, fostering dialogue and debate, and amplifying their voices to new audiences. It is to protect their voices -- and the voices of all those who use our service -- that we continue to work on the safety of the service.

Product features
The power of Twitter lies in the fact that we are an open, public and real time service. Our service is reflective of real conversations happening in the world and that sometimes includes perspectives that may be offensive, controversial, and/or bigoted to others.

We have a series of tools, built into our product, to help keep people safe and give them control over what they see and who they interact with. These tools include:

- **Unfollow**: If someone wants to stop seeing a particular account's Tweets in their home timeline, they can unfollow the account. They can still view the Tweets on an as-needed basis by visiting the profile, unless the Tweets on the profile are protected.
- **Block**: People can restrict specific accounts from contacting them, seeing their Tweets, and following them by blocking the account.
- **Advanced Block**: People can export their list of blocked accounts to share with another person and import someone else's list of blocked accounts using the Advanced Block feature.
- **Mute**: People can remove an account's Tweets from their timeline without unfollowing or blocking it. They can also use Advanced Mute for particular words, conversations, phrases, usernames, emojis, or hashtags.
- **Disable Receive Direct Message setting**: People can prevent accounts that they do not follow from DMing them by disabling the Receive Direct Message setting.



- **Filtered notifications:** People can apply different filters on the types of notifications they receive. Mute Notifications allows people to mute phrases and words they'd like to avoid seeing in their notifications. Advanced Filters allows them to disable notifications from certain types of accounts or at certain time periods - for example if their account is receiving a lot of sudden attention.
- **Protected Tweets:** When a person signs up for Twitter, their Tweets are public by default which means that anyone can view and interact with them. If a person protects their Tweets, this will make their account private and other Twitter users will have to send a request if they want to follow the account.
- **Safe search:** The Safe Search function removes potentially sensitive content by default, as well as accounts people have blocked and muted from search pages.
- **Sensitive media:** People can opt out of seeing certain imagery that may be sensitive. Twitter's default setting is to place potential sensitive material behind a warning. This can be adjusted in settings.

Updates to the Twitter Rules
The Twitter Rules are a living document and we are continually working to update, refine, and improve both our enforcement and our policies. This work is informed by in-depth research around trends in online behavior both on and off Twitter, feedback from the people who use Twitter, and input from a number of external entities.

Our rules are in place to ensure all people can participate in the public conversation freely and safely. Violence, harassment and other similar types of behavior discourage people from expressing themselves, and ultimately diminish the value of global public conversation.

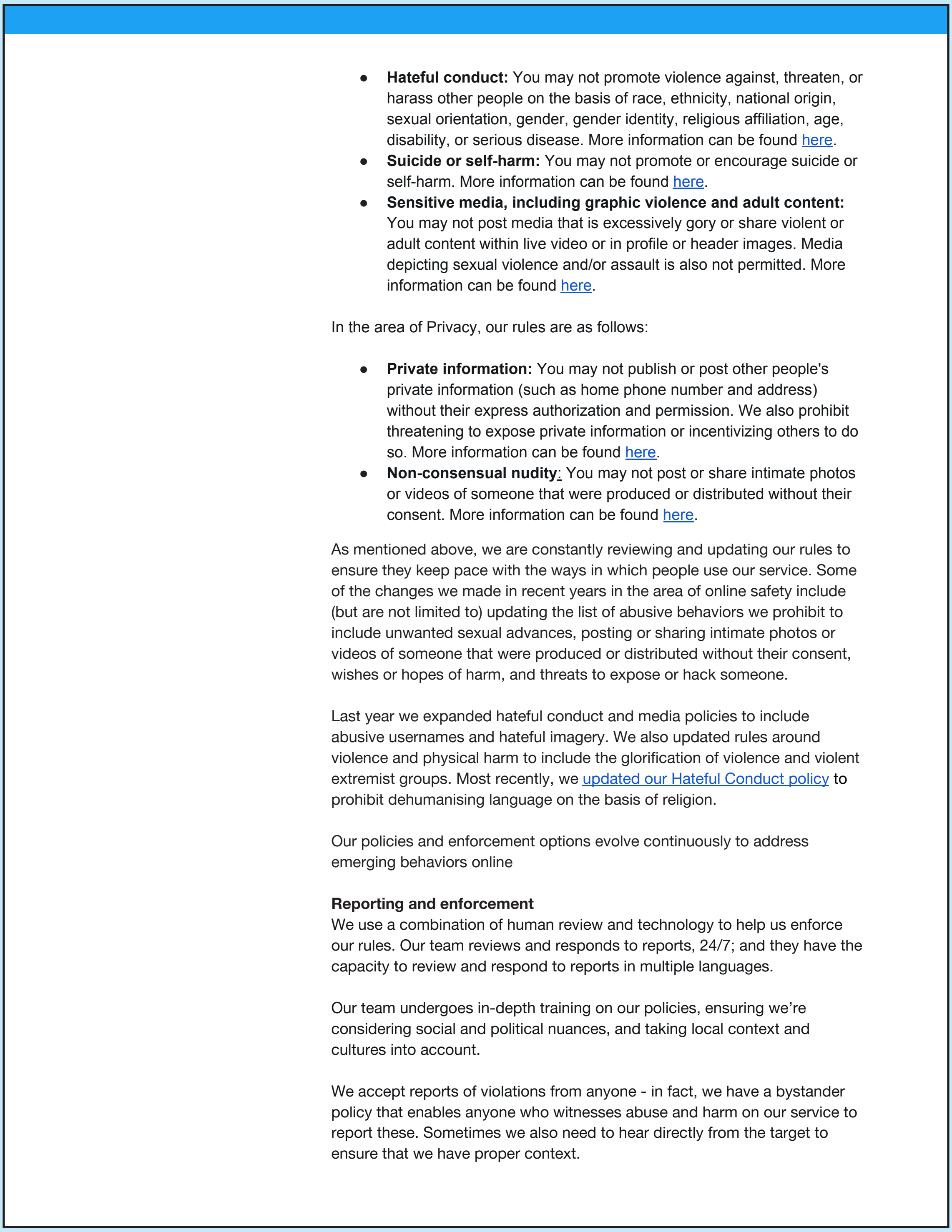
In June this year we undertook a major refresh of the Twitter Rules to make them simpler and easier to understand. We've gone from about 2,500 words to under 600. Each Rule is now 280 characters or less (the length of a Tweet) and describes exactly what is not allowed on Twitter.

We organised our rules around three categories — Safety, Privacy, and Authenticity — which makes it easier for people to find the information they're looking for more quickly.

Although we have simplified the language of our rules considerably, where possible, we've updated our rules pages to include more detail such as examples, step-by-step instructions about how to report, and details on what happens when we take action.

Relevant to your enquiry, in the area of Safety, our rules are as follows:

- **Violence:** You may not threaten violence against an individual or a group of people. We also prohibit the glorification of violence. Learn more about our violent threat and glorification of violence policies.
- **Child sexual exploitation:** We have zero tolerance for child sexual exploitation on Twitter. More information can be found here.
- **Abuse/harassment:** You may not engage in the targeted harassment of someone, or incite other people to do so. This includes wishing or hoping that someone experiences physical harm. More information can be found here.



- **Hateful conduct:** You may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. More information can be found here.
- **Suicide or self-harm:** You may not promote or encourage suicide or self-harm. More information can be found here.
- **Sensitive media, including graphic violence and adult content:** You may not post media that is excessively gory or share violent or adult content within live video or in profile or header images. Media depicting sexual violence and/or assault is also not permitted. More information can be found here.

In the area of Privacy, our rules are as follows:

- **Private information:** You may not publish or post other people's private information (such as home phone number and address) without their express authorization and permission. We also prohibit threatening to expose private information or incentivizing others to do so. More information can be found here.
- **Non-consensual nudity**: You may not post or share intimate photos or videos of someone that were produced or distributed without their consent. More information can be found here.

As mentioned above, we are constantly reviewing and updating our rules to ensure they keep pace with the ways in which people use our service. Some of the changes we made in recent years in the area of online safety include (but are not limited to) updating the list of abusive behaviors we prohibit to include unwanted sexual advances, posting or sharing intimate photos or videos of someone that were produced or distributed without their consent, wishes or hopes of harm, and threats to expose or hack someone.

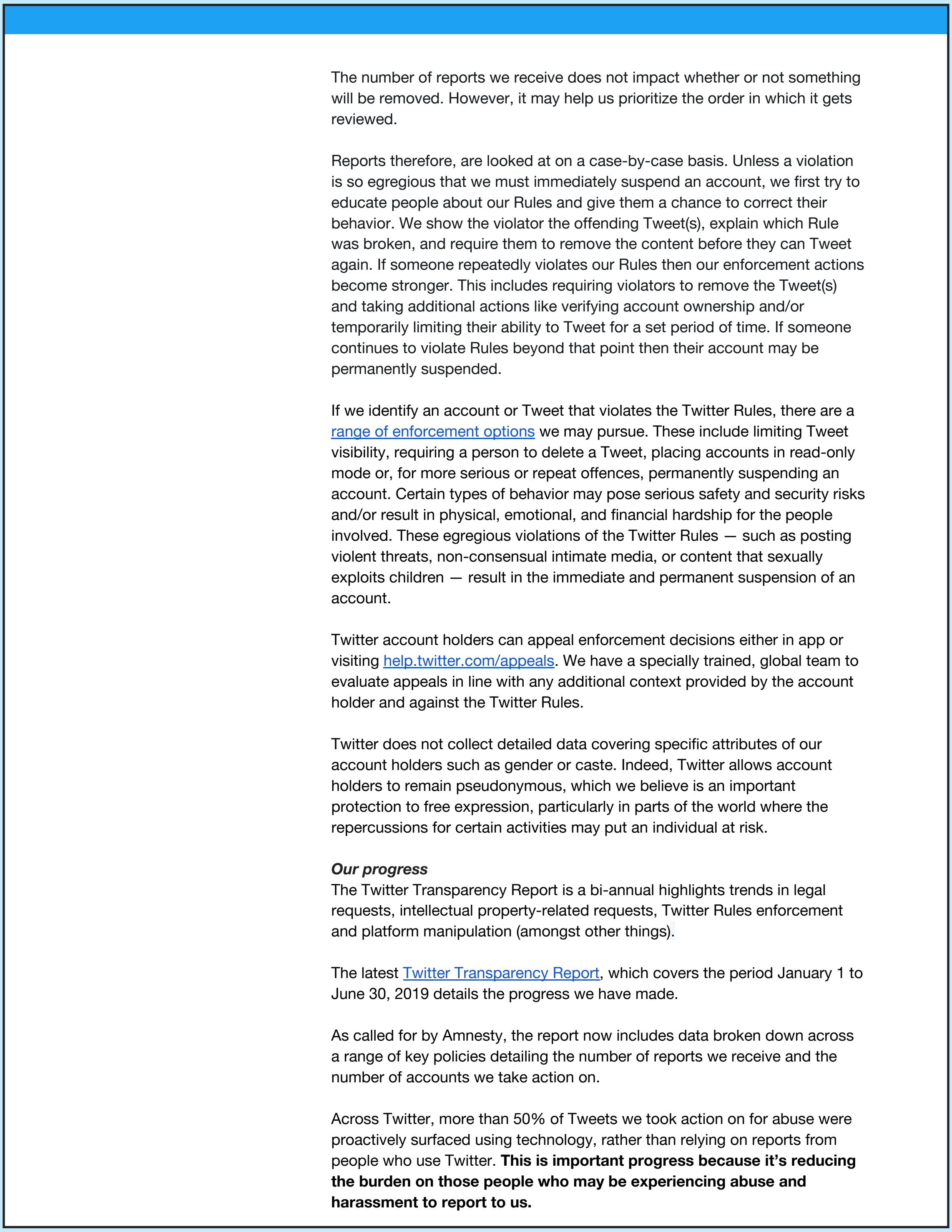
Last year we expanded hateful conduct and media policies to include abusive usernames and hateful imagery. We also updated rules around violence and physical harm to include the glorification of violence and violent extremist groups. Most recently, we updated our Hateful Conduct policy to prohibit dehumanising language on the basis of religion.

Our policies and enforcement options evolve continuously to address emerging behaviors online

Reporting and enforcement
We use a combination of human review and technology to help us enforce our rules. Our team reviews and responds to reports, 24/7; and they have the capacity to review and respond to reports in multiple languages.

Our team undergoes in-depth training on our policies, ensuring we’re considering social and political nuances, and taking local context and cultures into account.

We accept reports of violations from anyone - in fact, we have a bystander policy that enables anyone who witnesses abuse and harm on our service to report these. Sometimes we also need to hear directly from the target to ensure that we have proper context.



The number of reports we receive does not impact whether or not something will be removed. However, it may help us prioritize the order in which it gets reviewed.

Reports therefore, are looked at on a case-by-case basis. Unless a violation is so egregious that we must immediately suspend an account, we first try to educate people about our Rules and give them a chance to correct their behavior. We show the violator the offending Tweet(s), explain which Rule was broken, and require them to remove the content before they can Tweet again. If someone repeatedly violates our Rules then our enforcement actions become stronger. This includes requiring violators to remove the Tweet(s) and taking additional actions like verifying account ownership and/or temporarily limiting their ability to Tweet for a set period of time. If someone continues to violate Rules beyond that point then their account may be permanently suspended.

If we identify an account or Tweet that violates the Twitter Rules, there are a range of enforcement options we may pursue. These include limiting Tweet visibility, requiring a person to delete a Tweet, placing accounts in read-only mode or, for more serious or repeat offences, permanently suspending an account. Certain types of behavior may pose serious safety and security risks and/or result in physical, emotional, and financial hardship for the people involved. These egregious violations of the Twitter Rules — such as posting violent threats, non-consensual intimate media, or content that sexually exploits children — result in the immediate and permanent suspension of an account.

Twitter account holders can appeal enforcement decisions either in app or visiting help.twitter.com/appeals. We have a specially trained, global team to evaluate appeals in line with any additional context provided by the account holder and against the Twitter Rules.

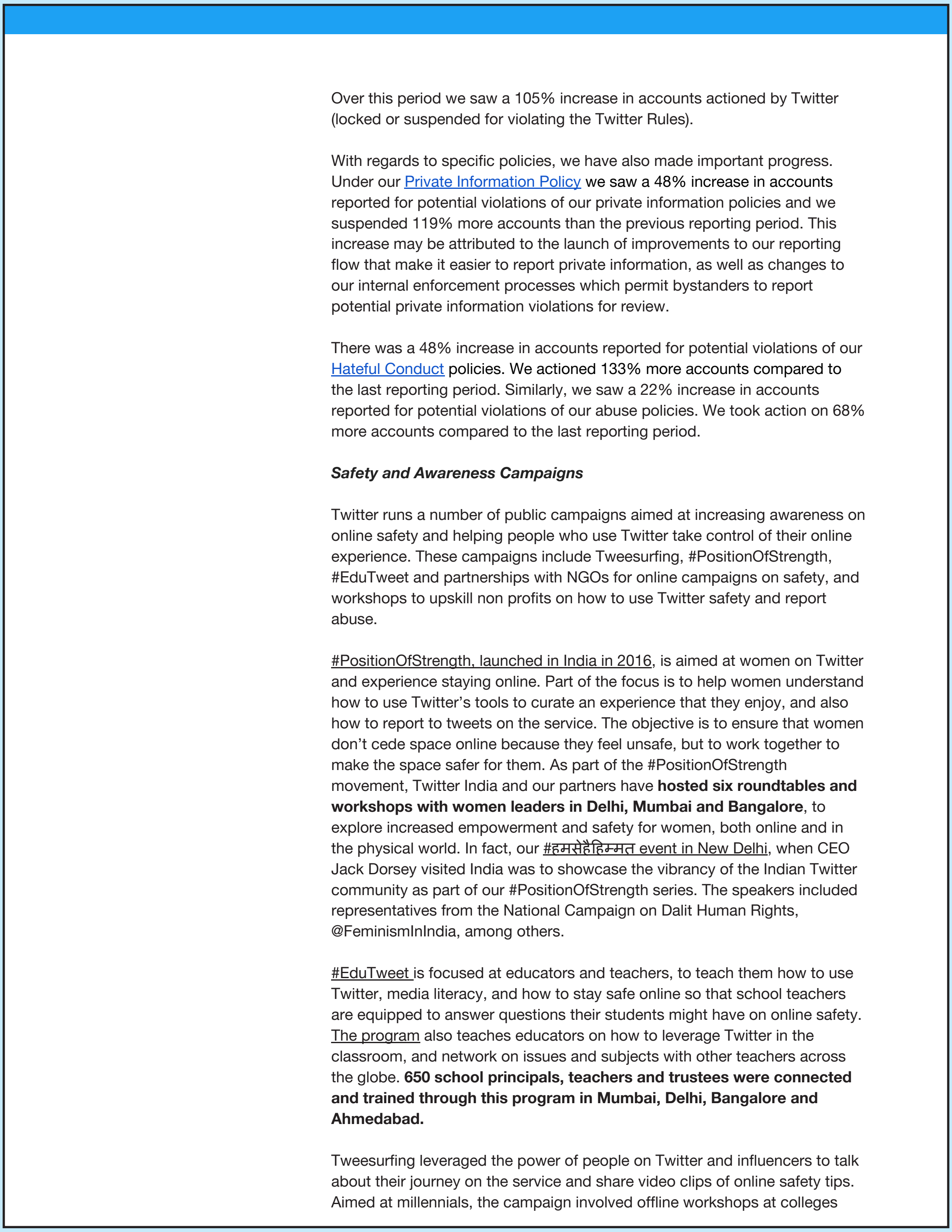
Twitter does not collect detailed data covering specific attributes of our account holders such as gender or caste. Indeed, Twitter allows account holders to remain pseudonymous, which we believe is an important protection to free expression, particularly in parts of the world where the repercussions for certain activities may put an individual at risk.

Our progress
The Twitter Transparency Report is a bi-annual highlights trends in legal requests, intellectual property-related requests, Twitter Rules enforcement and platform manipulation (amongst other things).

The latest Twitter Transparency Report, which covers the period January 1 to June 30, 2019 details the progress we have made.

As called for by Amnesty, the report now includes data broken down across a range of key policies detailing the number of reports we receive and the number of accounts we take action on.

Across Twitter, more than 50% of Tweets we took action on for abuse were proactively surfaced using technology, rather than relying on reports from people who use Twitter. **This is important progress because it's reducing the burden on those people who may be experiencing abuse and harassment to report to us.**



Over this period we saw a 105% increase in accounts actioned by Twitter (locked or suspended for violating the Twitter Rules).

With regards to specific policies, we have also made important progress. Under our Private Information Policy we saw a 48% increase in accounts reported for potential violations of our private information policies and we suspended 119% more accounts than the previous reporting period. This increase may be attributed to the launch of improvements to our reporting flow that make it easier to report private information, as well as changes to our internal enforcement processes which permit bystanders to report potential private information violations for review.

There was a 48% increase in accounts reported for potential violations of our Hateful Conduct policies. We actioned 133% more accounts compared to the last reporting period. Similarly, we saw a 22% increase in accounts reported for potential violations of our abuse policies. We took action on 68% more accounts compared to the last reporting period.

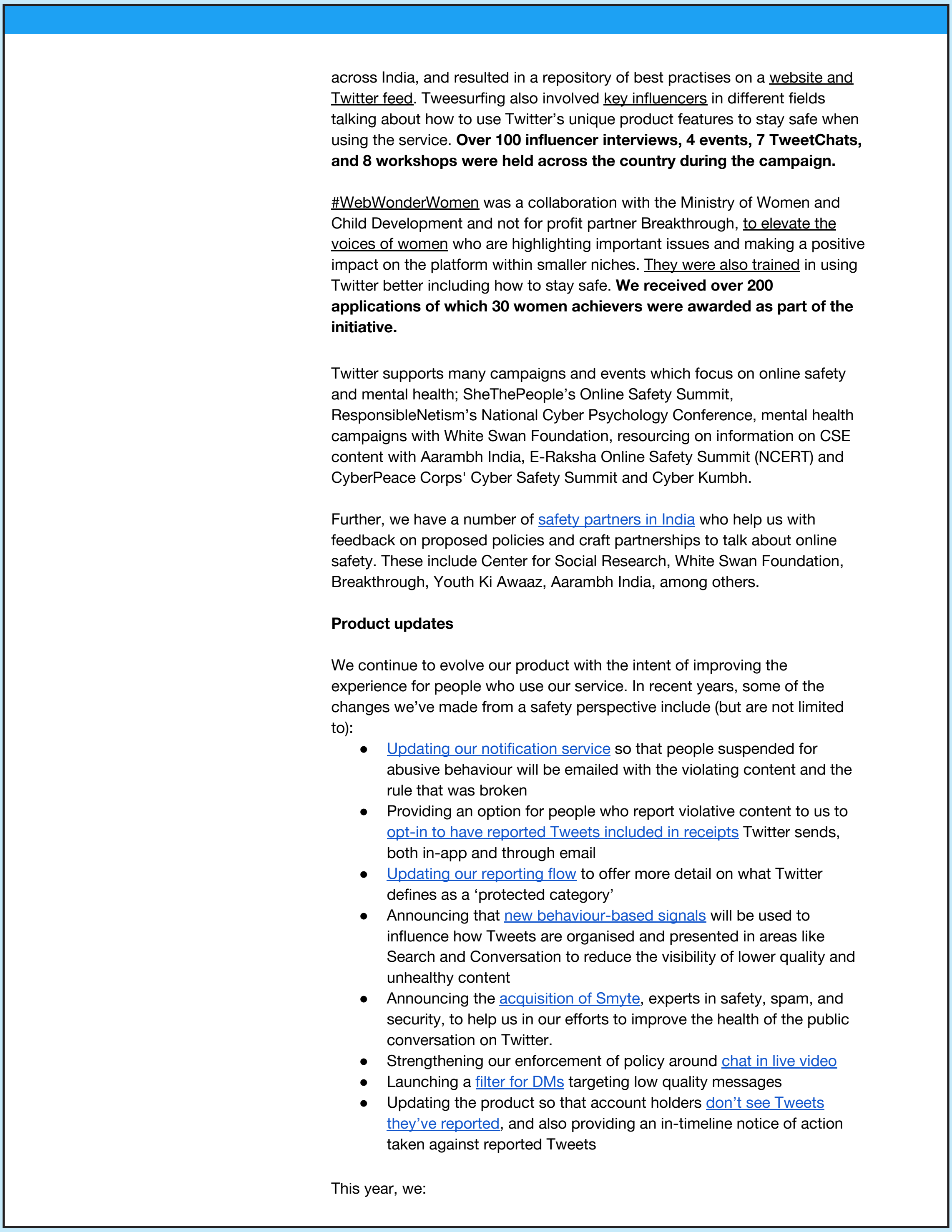
Safety and Awareness Campaigns

Twitter runs a number of public campaigns aimed at increasing awareness on online safety and helping people who use Twitter take control of their online experience. These campaigns include Tweesurfing, #PositionOfStrength, #EduTweet and partnerships with NGOs for online campaigns on safety, and workshops to upskill non profits on how to use Twitter safety and report abuse.

#PositionOfStrength, launched in India in 2016, is aimed at women on Twitter and experience staying online. Part of the focus is to help women understand how to use Twitter's tools to curate an experience that they enjoy, and also how to report to tweets on the service. The objective is to ensure that women don't cede space online because they feel unsafe, but to work together to make the space safer for them. As part of the #PositionOfStrength movement, Twitter India and our partners have **hosted six roundtables and workshops with women leaders in Delhi, Mumbai and Bangalore**, to explore increased empowerment and safety for women, both online and in the physical world. In fact, our #हमसेहैहिम्मत event in New Delhi, when CEO Jack Dorsey visited India was to showcase the vibrancy of the Indian Twitter community as part of our #PositionOfStrength series. The speakers included representatives from the National Campaign on Dalit Human Rights, @FeminismInIndia, among others.

#EduTweet is focused at educators and teachers, to teach them how to use Twitter, media literacy, and how to stay safe online so that school teachers are equipped to answer questions their students might have on online safety. The program also teaches educators on how to leverage Twitter in the classroom, and network on issues and subjects with other teachers across the globe. **650 school principals, teachers and trustees were connected and trained through this program in Mumbai, Delhi, Bangalore and Ahmedabad.**

Tweesurfing leveraged the power of people on Twitter and influencers to talk about their journey on the service and share video clips of online safety tips. Aimed at millennials, the campaign involved offline workshops at colleges



across India, and resulted in a repository of best practises on a website and Twitter feed. Tweesurfing also involved key influencers in different fields talking about how to use Twitter's unique product features to stay safe when using the service. **Over 100 influencer interviews, 4 events, 7 TweetChats, and 8 workshops were held across the country during the campaign.**

#WebWonderWomen was a collaboration with the Ministry of Women and Child Development and not for profit partner Breakthrough, to elevate the voices of women who are highlighting important issues and making a positive impact on the platform within smaller niches. They were also trained in using Twitter better including how to stay safe. **We received over 200 applications of which 30 women achievers were awarded as part of the initiative.**

Twitter supports many campaigns and events which focus on online safety and mental health; SheThePeople's Online Safety Summit, ResponsibleNetism's National Cyber Psychology Conference, mental health campaigns with White Swan Foundation, resourcing on information on CSE content with Aarambh India, E-Raksha Online Safety Summit (NCERT) and CyberPeace Corps' Cyber Safety Summit and Cyber Kumbh.

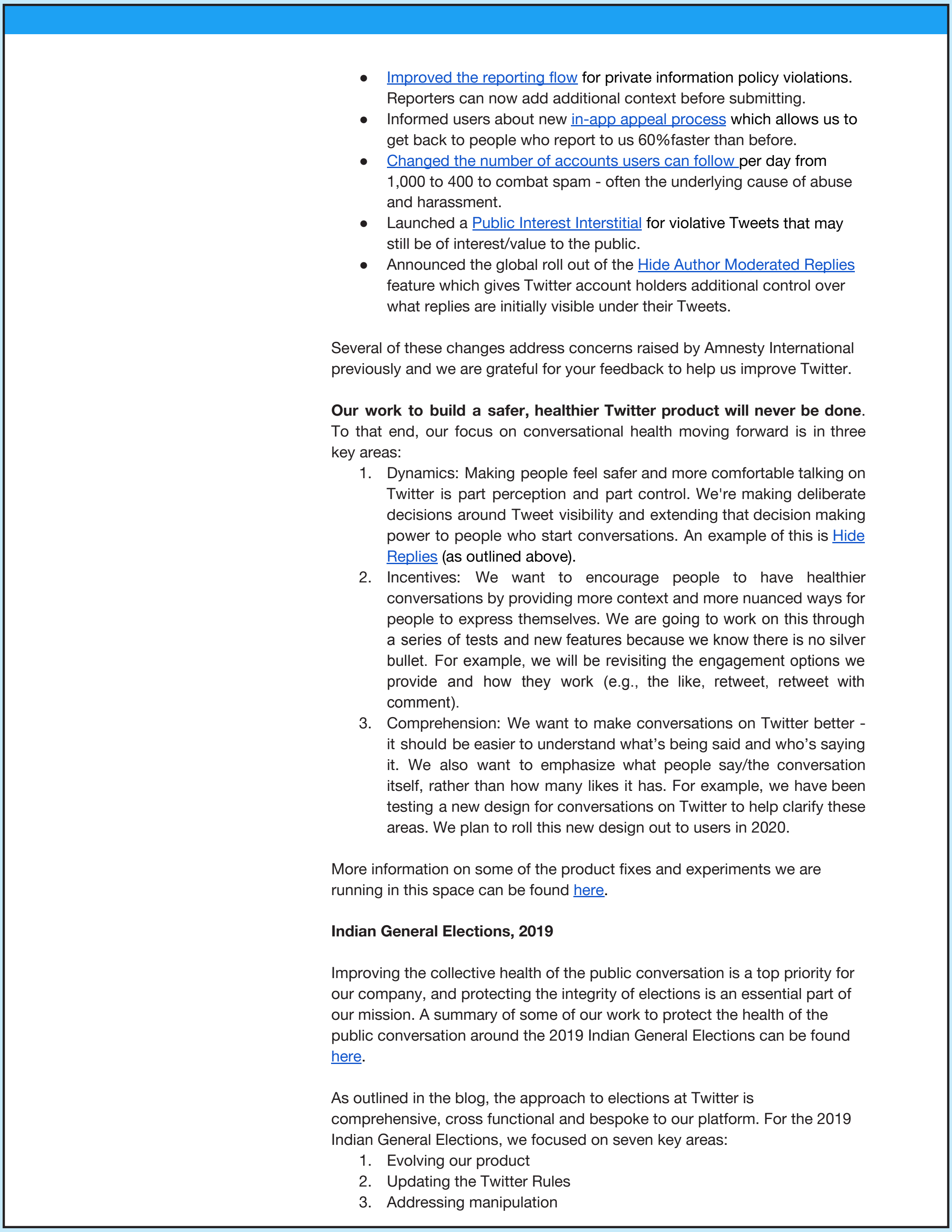
Further, we have a number of safety partners in India who help us with feedback on proposed policies and craft partnerships to talk about online safety. These include Center for Social Research, White Swan Foundation, Breakthrough, Youth Ki Awaaz, Aarambh India, among others.

Product updates

We continue to evolve our product with the intent of improving the experience for people who use our service. In recent years, some of the changes we've made from a safety perspective include (but are not limited to):

- Updating our notification service so that people suspended for abusive behaviour will be emailed with the violating content and the rule that was broken
- Providing an option for people who report violative content to us to opt-in to have reported Tweets included in receipts Twitter sends, both in-app and through email
- Updating our reporting flow to offer more detail on what Twitter defines as a 'protected category'
- Announcing that new behaviour-based signals will be used to influence how Tweets are organised and presented in areas like Search and Conversation to reduce the visibility of lower quality and unhealthy content
- Announcing the acquisition of Smyte, experts in safety, spam, and security, to help us in our efforts to improve the health of the public conversation on Twitter.
- Strengthening our enforcement of policy around chat in live video
- Launching a filter for DMs targeting low quality messages
- Updating the product so that account holders don't see Tweets they've reported, and also providing an in-timeline notice of action taken against reported Tweets

This year, we:



- Improved the reporting flow for private information policy violations. Reporters can now add additional context before submitting.
- Informed users about new in-app appeal process which allows us to get back to people who report to us 60%faster than before.
- Changed the number of accounts users can follow per day from 1,000 to 400 to combat spam - often the underlying cause of abuse and harassment.
- Launched a Public Interest Interstitial for violative Tweets that may still be of interest/value to the public.
- Announced the global roll out of the Hide Author Moderated Replies feature which gives Twitter account holders additional control over what replies are initially visible under their Tweets.

Several of these changes address concerns raised by Amnesty International previously and we are grateful for your feedback to help us improve Twitter.

Our work to build a safer, healthier Twitter product will never be done. To that end, our focus on conversational health moving forward is in three key areas:

1. Dynamics: Making people feel safer and more comfortable talking on Twitter is part perception and part control. We're making deliberate decisions around Tweet visibility and extending that decision making power to people who start conversations. An example of this is Hide Replies (as outlined above).
2. Incentives: We want to encourage people to have healthier conversations by providing more context and more nuanced ways for people to express themselves. We are going to work on this through a series of tests and new features because we know there is no silver bullet. For example, we will be revisiting the engagement options we provide and how they work (e.g., the like, retweet, retweet with comment).
3. Comprehension: We want to make conversations on Twitter better - it should be easier to understand what's being said and who's saying it. We also want to emphasize what people say/the conversation itself, rather than how many likes it has. For example, we have been testing a new design for conversations on Twitter to help clarify these areas. We plan to roll this new design out to users in 2020.

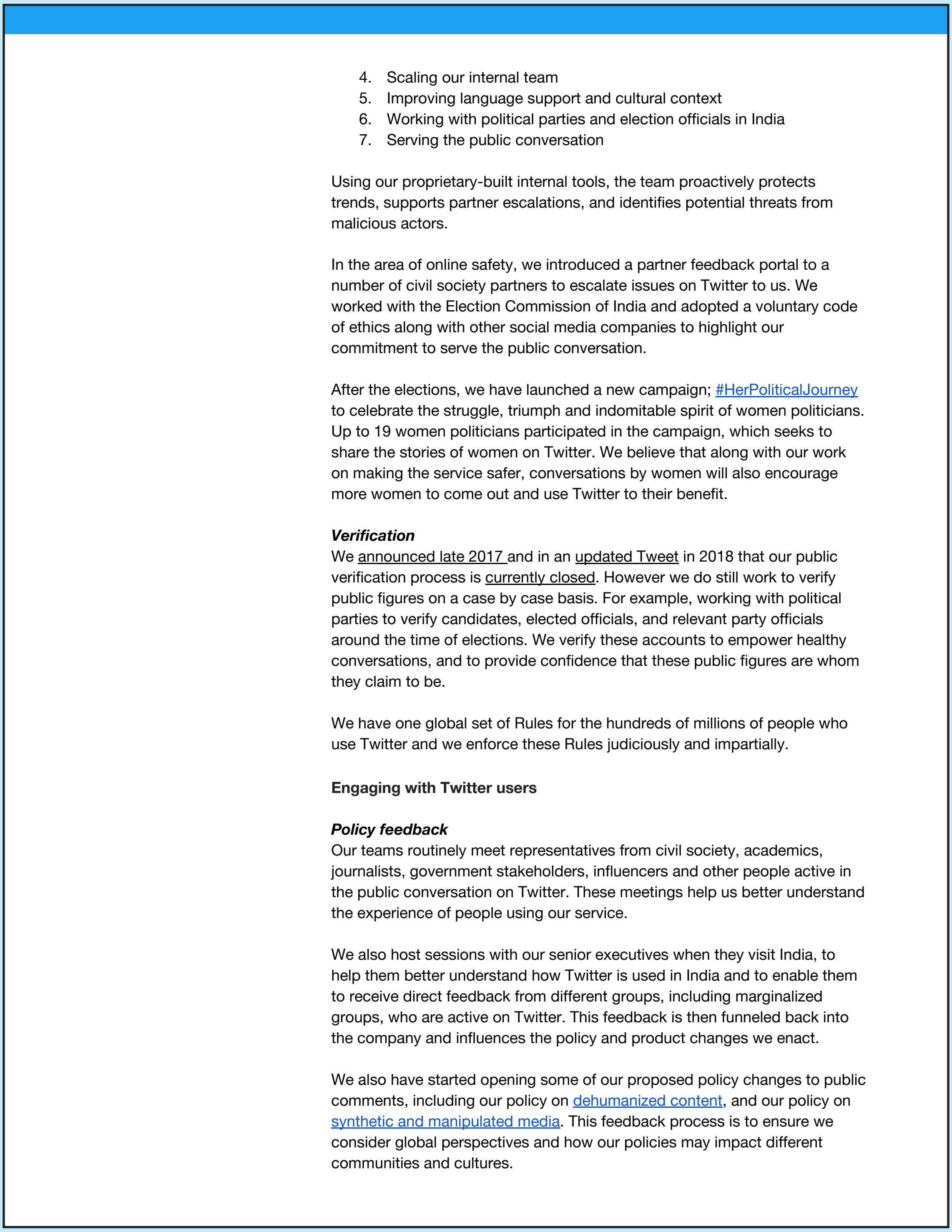
More information on some of the product fixes and experiments we are running in this space can be found here.

Indian General Elections, 2019

Improving the collective health of the public conversation is a top priority for our company, and protecting the integrity of elections is an essential part of our mission. A summary of some of our work to protect the health of the public conversation around the 2019 Indian General Elections can be found here.

As outlined in the blog, the approach to elections at Twitter is comprehensive, cross functional and bespoke to our platform. For the 2019 Indian General Elections, we focused on seven key areas:

1. Evolving our product
2. Updating the Twitter Rules
3. Addressing manipulation



4. Scaling our internal team
5. Improving language support and cultural context
6. Working with political parties and election officials in India
7. Serving the public conversation

Using our proprietary-built internal tools, the team proactively protects trends, supports partner escalations, and identifies potential threats from malicious actors.

In the area of online safety, we introduced a partner feedback portal to a number of civil society partners to escalate issues on Twitter to us. We worked with the Election Commission of India and adopted a voluntary code of ethics along with other social media companies to highlight our commitment to serve the public conversation.

After the elections, we have launched a new campaign; #HerPoliticalJourney to celebrate the struggle, triumph and indomitable spirit of women politicians. Up to 19 women politicians participated in the campaign, which seeks to share the stories of women on Twitter. We believe that along with our work on making the service safer, conversations by women will also encourage more women to come out and use Twitter to their benefit.

Verification
We announced late 2017 and in an updated Tweet in 2018 that our public verification process is currently closed. However we do still work to verify public figures on a case by case basis. For example, working with political parties to verify candidates, elected officials, and relevant party officials around the time of elections. We verify these accounts to empower healthy conversations, and to provide confidence that these public figures are whom they claim to be.

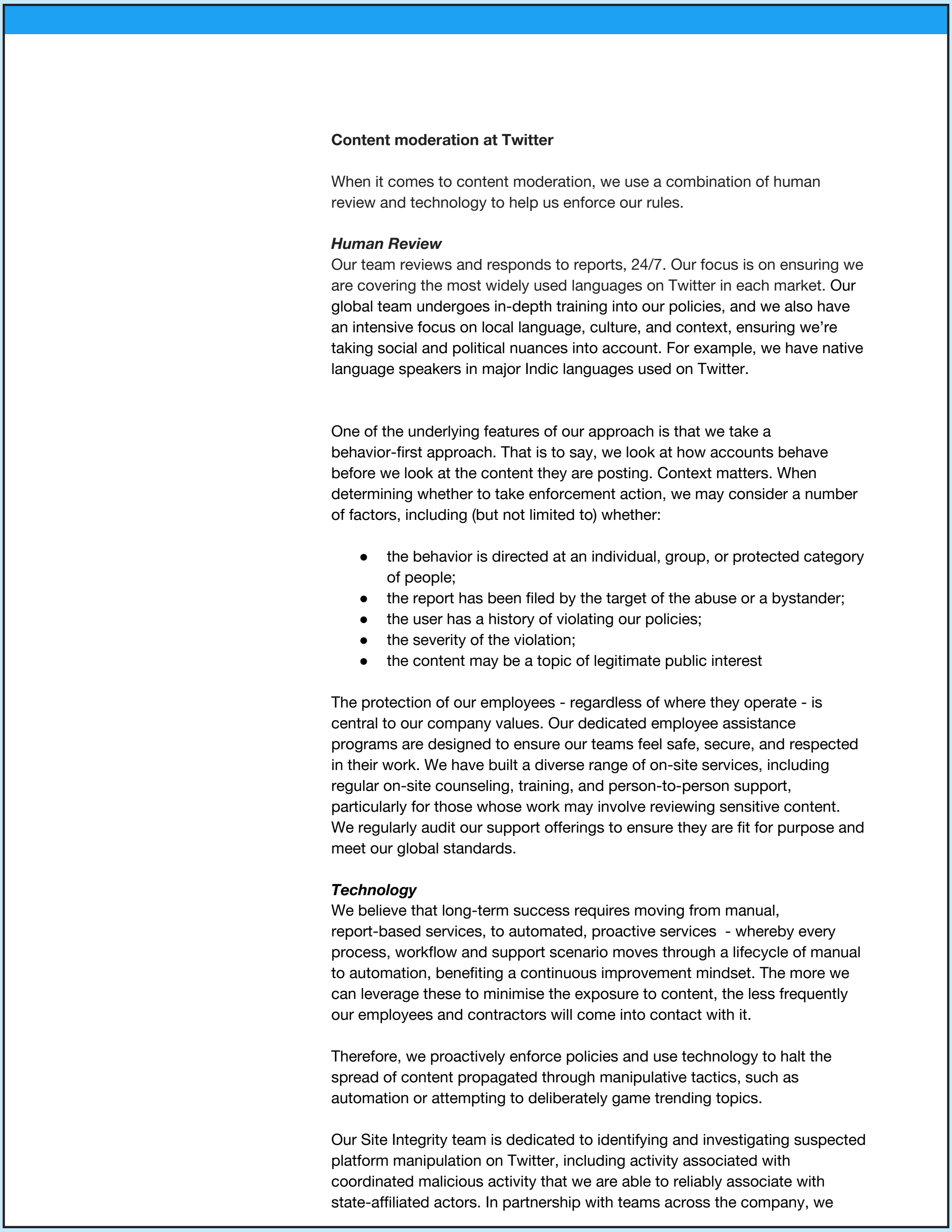
We have one global set of Rules for the hundreds of millions of people who use Twitter and we enforce these Rules judiciously and impartially.

Engaging with Twitter users

Policy feedback
Our teams routinely meet representatives from civil society, academics, journalists, government stakeholders, influencers and other people active in the public conversation on Twitter. These meetings help us better understand the experience of people using our service.

We also host sessions with our senior executives when they visit India, to help them better understand how Twitter is used in India and to enable them to receive direct feedback from different groups, including marginalized groups, who are active on Twitter. This feedback is then funneled back into the company and influences the policy and product changes we enact.

We also have started opening some of our proposed policy changes to public comments, including our policy on dehumanized content, and our policy on synthetic and manipulated media. This feedback process is to ensure we consider global perspectives and how our policies may impact different communities and cultures.



Content moderation at Twitter

When it comes to content moderation, we use a combination of human review and technology to help us enforce our rules.

Human Review
Our team reviews and responds to reports, 24/7. Our focus is on ensuring we are covering the most widely used languages on Twitter in each market. Our global team undergoes in-depth training into our policies, and we also have an intensive focus on local language, culture, and context, ensuring we're taking social and political nuances into account. For example, we have native language speakers in major Indic languages used on Twitter.

One of the underlying features of our approach is that we take a behavior-first approach. That is to say, we look at how accounts behave before we look at the content they are posting. Context matters. When determining whether to take enforcement action, we may consider a number of factors, including (but not limited to) whether:

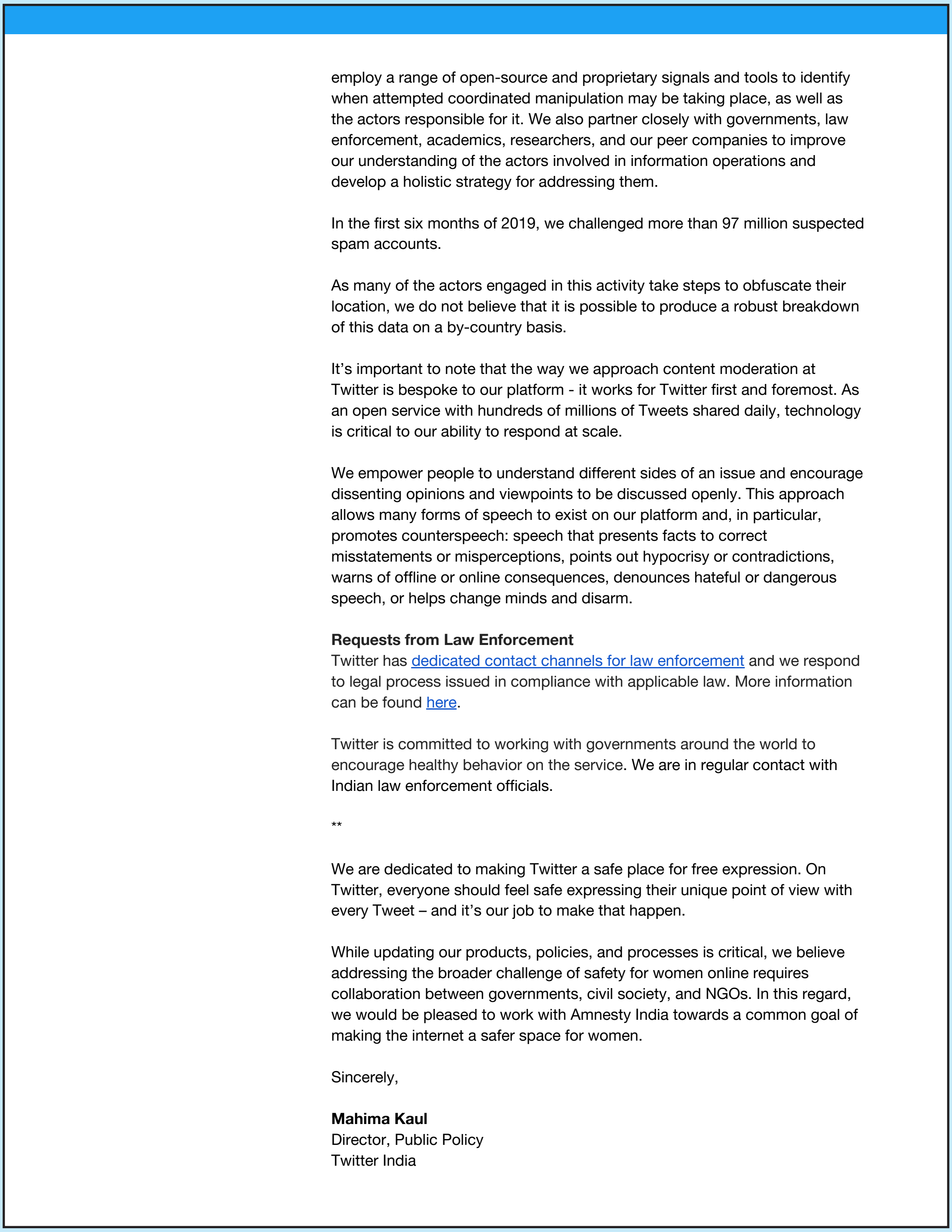
- the behavior is directed at an individual, group, or protected category of people;
- the report has been filed by the target of the abuse or a bystander;
- the user has a history of violating our policies;
- the severity of the violation;
- the content may be a topic of legitimate public interest

The protection of our employees - regardless of where they operate - is central to our company values. Our dedicated employee assistance programs are designed to ensure our teams feel safe, secure, and respected in their work. We have built a diverse range of on-site services, including regular on-site counseling, training, and person-to-person support, particularly for those whose work may involve reviewing sensitive content. We regularly audit our support offerings to ensure they are fit for purpose and meet our global standards.

Technology
We believe that long-term success requires moving from manual, report-based services, to automated, proactive services - whereby every process, workflow and support scenario moves through a lifecycle of manual to automation, benefiting a continuous improvement mindset. The more we can leverage these to minimise the exposure to content, the less frequently our employees and contractors will come into contact with it.

Therefore, we proactively enforce policies and use technology to halt the spread of content propagated through manipulative tactics, such as automation or attempting to deliberately game trending topics.

Our Site Integrity team is dedicated to identifying and investigating suspected platform manipulation on Twitter, including activity associated with coordinated malicious activity that we are able to reliably associate with state-affiliated actors. In partnership with teams across the company, we



employ a range of open-source and proprietary signals and tools to identify when attempted coordinated manipulation may be taking place, as well as the actors responsible for it. We also partner closely with governments, law enforcement, academics, researchers, and our peer companies to improve our understanding of the actors involved in information operations and develop a holistic strategy for addressing them.

In the first six months of 2019, we challenged more than 97 million suspected spam accounts.

As many of the actors engaged in this activity take steps to obfuscate their location, we do not believe that it is possible to produce a robust breakdown of this data on a by-country basis.

It's important to note that the way we approach content moderation at Twitter is bespoke to our platform - it works for Twitter first and foremost. As an open service with hundreds of millions of Tweets shared daily, technology is critical to our ability to respond at scale.

We empower people to understand different sides of an issue and encourage dissenting opinions and viewpoints to be discussed openly. This approach allows many forms of speech to exist on our platform and, in particular, promotes counterspeech: speech that presents facts to correct misstatements or misperceptions, points out hypocrisy or contradictions, warns of offline or online consequences, denounces hateful or dangerous speech, or helps change minds and disarm.

Requests from Law Enforcement
Twitter has dedicated contact channels for law enforcement and we respond to legal process issued in compliance with applicable law. More information can be found here.

Twitter is committed to working with governments around the world to encourage healthy behavior on the service. We are in regular contact with Indian law enforcement officials.

**

We are dedicated to making Twitter a safe place for free expression. On Twitter, everyone should feel safe expressing their unique point of view with every Tweet – and it's our job to make that happen.

While updating our products, policies, and processes is critical, we believe addressing the broader challenge of safety for women online requires collaboration between governments, civil society, and NGOs. In this regard, we would be pleased to work with Amnesty India towards a common goal of making the internet a safer space for women.

Sincerely,

Mahima Kaul
Director, Public Policy
Twitter India

INDIA AMNESTY INTERNATIONAL

Indians For Amnesty International Trust
#235, Ground Floor, 13th Cross, Indira Nagar,
2nd Stage, Bangalore – 560038, Karnataka, India
www.amnesty.org.in

CONTACT US

✉ contact@amnesty.org.in

☎ +91 80 49388000

JOIN THE CONVERSATION

www.facebook.com/AIIndia

@AIIndia