

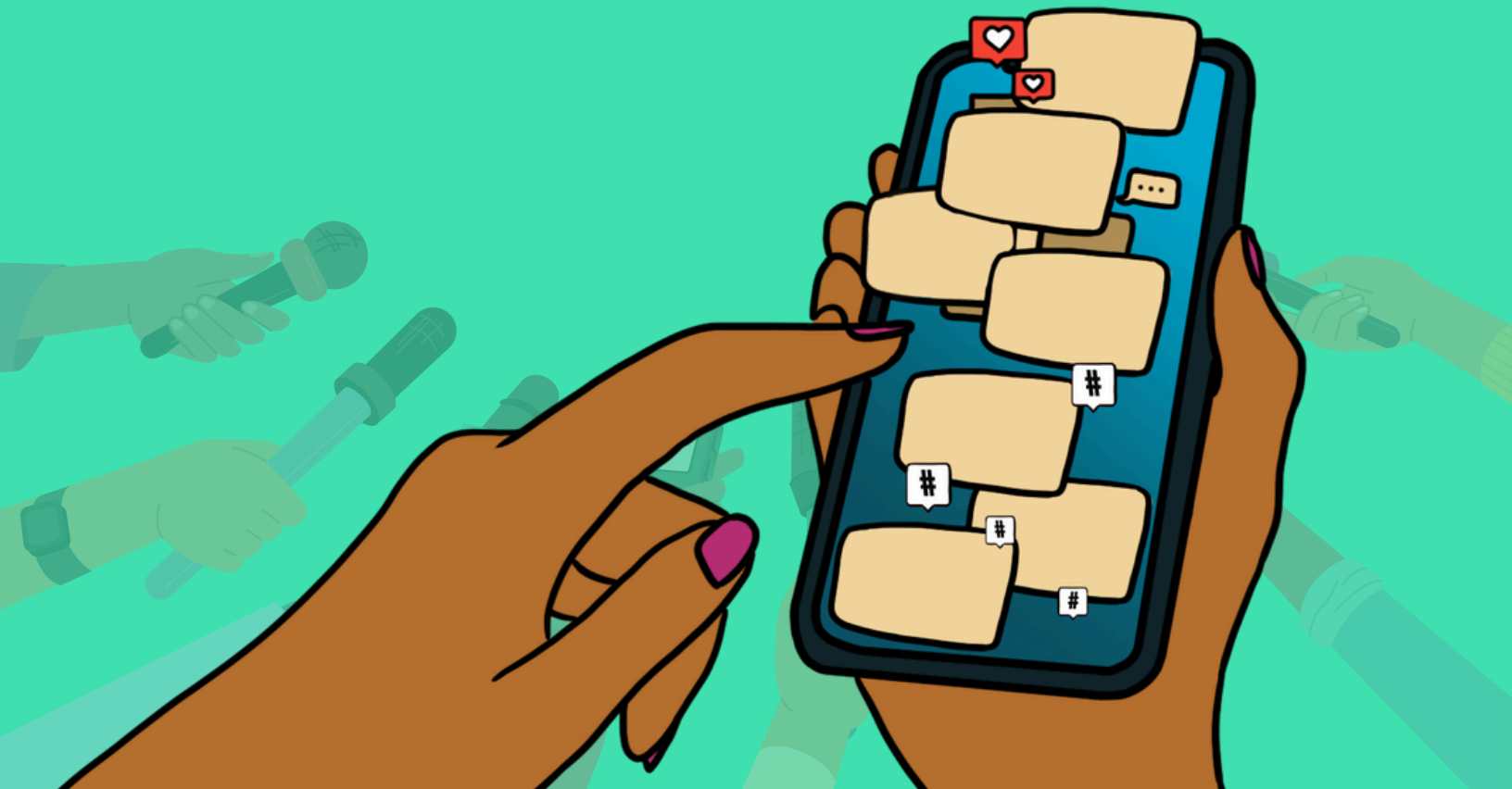


DigitalRightsFoundation
"KNOW YOUR RIGHTS"

GENDERED

ONLINE HATE IN PAKISTAN

*RIGHT-WING RELIGIOUS CAMPAIGNS AGAINST
WOMEN JOURNALISTS*



© Digital Rights Foundation

December 2024

Researched and co-authored by: Aleena Afzaal, Maryam Ali Khan, Ashus Owaisi, and Abdullah Tariq.

Edited by: Adnan Chaudhri, Seerat Khan, and Talal Raza.

Design and layout by: Ahsan Zahid and Talha Umar.

Digital Rights Foundation (DRF) is a women-led, not-for-profit organisation based in Pakistan working on digital rights freedoms since 2013. DRF envisions a place where all people, especially women and gender minorities, can exercise the right of expression without being threatened. DRF believes that a free internet with access to information and impeccable privacy policies can create safe online spaces for not only women but the world at large.

At DRF, we aim to strengthen the protections for human rights defenders (HRDs), with a focus on women's rights in digital spaces through policy advocacy and digital security awareness-raising. In addition, one of our main aims is to protect women from cyber harassment that they have to deal with throughout their lives by making them aware of their rights and making resources accessible when they need help.

With growing privacy concerns in digital spaces, DRF seeks to increase awareness about privacy issues and defend the right to privacy through research, monitoring and reporting the tactics around surveillance. To rally with other actors for strong legal protection for privacy in the country and to raise our voices against dictating censorship policies, we propose viable solutions to the government and other authoritative bodies.

Contact information:

info@digitalrightsfoundation.pk

www.digitalrightsfoundation.pk

TABLE OF CONTENTS

01	Executive Summary	1
02	Introduction	2
03	The Impact of Censorship and Pakistan's Blasphemy Laws	5
04	Party Y & Their Online Political Attack on Women Journalist Z	7
05	Theoretical Framework	9
06	Methodology	10
07	Findings & Analysis	14
	<ul style="list-style-type: none">• Chi-square Correlation Test• Contingency Table for 'Sectarian & Communal Hatred' and 'Religious Intolerance'• Contingency Table for 'Gender Based Hatred' and 'Derogatory Language & Name-Calling'• Contingency Table for 'Revolutionary' and 'Mobilising Campaigns'	

- Contingency Table for 'Gender Based Hatred' and 'Anti-Authority Sentiments'
- Pairwise Support Analysis Heatmap
- Role of X in amplifying hateful content in Pakistan Impact on Journalist Z

08 Platform Policies: Addressing Gendered Hate Speech and Violent Content 28

09 Recommendations 32

Disclaimer: *The woman journalist and the political organization's identities have been anonymized due to ethical research considerations. The woman journalist will thereby be referred to as "Z" while the political organization will be referred to as "Party Y".*

Executive Summary

Women journalists are on average subjected to significantly higher online harassment, particularly in the form of gendered hate speech, than their male counterparts. In May 2024, a Pakistani woman journalist raised her voice on X (formerly Twitter) against a case of mob violence that turned fatal, and that arose as a consequence of allegations of blasphemy. When the journalist in question stated that a radical, far right religious group in the country was behind the act, this resulted in intense backlash from the group's supporters and workers on the social media platform. In this case study, we explore theoretical frameworks anchored in the public sphere and gender biases leading to eventual hatred. Data collected from X (the journalist's tweets, user interactions, quoted tweets, hashtags, mentions, associated keywords, etc.) was used to analyze the intensity of the backlash that the woman journalist was subjected to, by defining multiple variables and categories that the online hate speech falls under. Our analysis of 565 tweets that she received in response - over 300 of which addressed her directly - highlighted "Gender Based Hatred", "Derogatory Language & Name-Calling" and "Threatening Language", with overlaps with "Sectarian & Communal Hatred" and "Mobilising Campaigns", among others themes.

We compared the data to a male journalist's tweets on the incident and the responses he received in order to understand the difference in public opinion through a gendered lens. The data was analyzed using correlation and association rules mining techniques such as chi square tests, *Apriori* analysis and other forms of data observation. We provide

recommendations to the platform to resolve content moderation issues that have been highlighted in the study. The study's observations show that online hate speech, particularly against women journalists, is not only disproportionately higher but also life threatening. Under such grave circumstances, social media platforms play a key role in ensuring user safety in online spaces.

Introduction

Media censorship in Pakistan has a long and complex history, shaped by the country's tumultuous political landscape and often characterized by state control and suppression of free speech, going back to the nation's creation in 1947. Pakistan's press has faced continuous series of restrictions, especially during one of Pakistan's numerous periods of military rule, where censorship was wielded as a tool to control narratives and limit dissenting voices.¹ During General Ayub Khan's dictatorship in the 1960s, for example, the government introduced the *Press and Publications Ordinance*, which gave authorities sweeping powers to ban publications that were deemed harmful to national security or ideology. These restrictions deepened under the regime of General Zia-ul-Haq (1978-1988), during which media censorship was institutionalised to support his Islamization policies.²

General Zia's era saw journalists subjected to intense scrutiny and censorship, and press freedom severely curtailed. Newspapers were required to submit their content for government approval, with stories critical of the state, or particular religious ideologies, routinely redacted. Journalists who defied these restrictions faced imprisonment, physical abuse, and even death.³ This legacy of censorship continued under subsequent governments, and even during periods of

civilian rule, Pakistan's media has remained under pressure. For instance, during the tenure of Prime Minister Nawaz Sharif (1997-1999) and later under General Pervez Musharraf (1999-2008), media organizations faced increased restrictions from government authorities, while autonomous regulatory bodies such as PEMRA (Pakistan Electronic Media Regulatory Authority) were used to suppress critical reporting.⁴

The media landscape in Pakistan continues to be marred by censorship, with political stakeholders using both overt and covert methods to influence media content. According to the World Press Freedom Index, Pakistan, at the time of this report, has a cumulative score of 33.9 based on various press freedom indexes – ranking Pakistan 152/180 in 2024.⁵ At least 6 journalists have been killed in 2024 alone, and one journalist imprisoned for “endangering public safety” through their reporting.⁶ Journalists are often pressured to practise self-censorship to avoid retaliation from powerful state and non-state actors, including the military and radical religious groups.⁷ In recent years, this censorship has extended to digital media. Social media platforms such as X (formerly Twitter) and Facebook have provided an alternative space for marginalized Pakistani voices, but the government of Pakistan has aggressively pursued tighter control over online content. Laws such as the *Prevention of Electronic Crimes Act* (PECA) allow authorities to monitor and censor online

material, often under the guise of national security or religious protection.⁸

In addition to restrictive legal and executive measures pertaining to digital media, social media platforms have become breeding grounds for orchestrated and coordinated attacks against various political actors. As part of its elections monitoring in regards to Pakistan's February 2024 general elections, DRF collected a sample of 225 posts, using online data collection tools to observe and break down how TFGBV and gendered disinformation – material against electoral candidates, activists and journalists was being disseminated. Of all the collected posts, 179 were collected from X, 17 from Instagram, 12 from Facebook, 9 from TikTok, 3 from YouTube while 5 were directly distributed. Furthermore, out of 225 posts, 163 fell under the category of gendered disinformation. Among PML-N, PTI and PPP, the highest number of gendered disinformation posts at 76 were targeted at PML-N, particularly related to Maryam Nawaz, followed by PTI at 61 and PPP at 19. While all platforms have their share in exacerbating the problem, X's (formerly Twitter) reduced human oversight in content moderation, and greater reliance on automated systems⁹ in particular, have resulted in the platform (along with others), becoming a more effective tool for coordinated attacks against women journalists. According to a study by the International Centre for Journalists, Facebook and X were rated as the two

least safe platforms by women journalists – at 39% and 26% respectively.¹⁰ AI-based content moderation tools are, in theory, efficient in handling large volumes of data. In reality, however, said tools are often unable to detect the relatively nuanced nature of gender-based abuse, allowing harmful content to persist on the platform.¹¹ Further to this, X's algorithmic focus on engagement further aggravates the issue by prioritizing sensational content, which will often include harmful and abusive messages. This shift has amplified the visibility of offensive content targeting women journalists, creating an online environment where gendered abuse is not only prevalent but often goes unaddressed.¹² These algorithmic content moderation practices on platforms unintentionally encourage further growth of organised hate campaigns, where coordinated efforts to silence dissenting voices are ramped up. As a consequence, women journalists are left vulnerable to coordinated online harassment, often politically driven, where troll armies and organized groups use misogynistic abuse to undermine their professional credibility.¹³

Women journalists and human rights defenders in Pakistan in particular face gender-based attacks, subjected to harassment and surveillance owing to their gender, at rates that usually outpace that of their male counterparts. In this case study, Journalist Z experienced similar abuse when criticising Party Y, becoming the target of an orchestrated online attack involving organised online trolls and

other groups spreading misogynistic abuse. As a study by the International Center for Journalists (ICJ) highlights that approximately 73 percent of women journalists face online violence, including threats of physical and sexual harm, abusive language, and digital security attacks.¹⁴ Separately, a study by the Digital Rights Foundation (DRF) found that women journalists were not only vulnerable to government surveillance but also to social surveillance by their audiences, political parties, non-state actors, as well as fellow journalists, and personal contacts.¹⁵ The convergence of media censorship, religious conservatism, and blasphemy laws creates a hostile environment for female journalists. The state's complex relationship with these religious extremist groups allows for an environment where criticism of religious or political figures can lead to severe consequences, including legal action under blasphemy laws.¹⁶

These unprecedented attacks by state and non-state actors have a profound impact on women's mental health, career choices, and overall well-being. Studies, such as those conducted by the EU DisinfoLab, reveal that these campaigns are often part of larger disinformation efforts. They are deliberately designed to exploit social media algorithms, increasing the visibility and spread of hate propaganda.¹⁷ Gharida Farooqi, a prominent television anchor, has been the target of multiple online harassment campaigns. In January 2024, she faced a

surge in threats and abusive messages, including rape threats and character assassination, as part of a coordinated effort to intimidate her.¹⁸ Similarly, Asma Shirazi, recognized for her journalism, has been subjected to online abuse and disinformation campaigns attempting to discredit her work.¹⁹ In July 2024, Dr. Omer Adil, a medical professional and former chairman of the Punjab Healthcare Commission (PHC), passed sexist and derogatory remarks against Gharida Farooqi and women in the media. His comments were seen as part of the larger issue of gender-based discrimination and harassment faced by women journalists in Pakistan. In response to his statements, over 60 journalists and media practitioners across Pakistan issued a joint statement with DRF, condemning his remarks and calling for accountability.^{20 21}

Through this case study, DRF aims to highlight current gaps in existing legal frameworks that fail to adequately protect individuals, primarily women journalists, from online harassment and hate speech; as well as the lack of policy reforms that aim to further silence victims and empower perpetrators in online spaces, in spite of numerous collective calls by women journalists, human rights defenders and others highlighting the inaction on the part of large social media platforms.

The Impact of Censorship and Pakistan's Blasphemy Laws

Pakistan's history of censorship and free speech restrictions have adversely affected citizens' fundamental freedoms in the country. The country's history has been marked by the misuse of blasphemy laws, often targeting marginalised groups particularly religious minorities. To this date, Pakistan holds one of the strictest sets of blasphemy laws globally, second only to Iran. These laws criminalise any insult toward the Prophet Muhammad and his companions, with penalties ranging from fines to the death sentence. However, the vague definitions and wide scope of Pakistan Penal Code (PPC) and PECA have led to rampant misuse. For example, Section 295-A of the PPC vaguely criminalizes acts that outrage religious feelings, without clear guidelines on what constitutes "deliberate and malicious" intent. Similarly, Section 295-C mandates the death penalty for derogatory remarks about the Prophet Muhammad but fails to define intent or allow for repentance. Sections 298-A, B & C penalizes speech against holy figures and restrict Ahmadi religious practices, but lack clarity, making them open to misuse. Section 11 of PECA extends these ambiguities online, criminalizing hate

speech with no procedural safeguards, often leading to wrongful charges.²² It has been noted by many that accusations of blasphemy are weaponised to settle personal scores, target religious minorities, and stifle political dissent. Extrajudicial killings and mob violence are frequent outcomes, as seen in high-profile cases like that of Asia Bibi, a Christian woman who spent years on death row before being acquitted which led to mob violence across the country.²³

Numerous cases have been reported relating to scurrilous blasphemy accusations leading to mob violence. In August 2023, two Christian men from the city of Jaranwala were accused on social media of allegedly desecrating a copy of the Quran, with torn and vandalised pages being shared as alleged "evidence." The collective furore that erupted, with Muslim clerics inciting violence against both the men and their community at large without evidence, resulted in churches being burned down or vandalised, and the homes of Christians of Jaranwala also attacked. Members of the city's Christian community fled to avoid the violence, while others were protected by their Muslim neighbours.²⁴ Though several people were arrested in the wake of the mob violence, have been realised, with no hope on the part of the Christians of Jaranwala of legal action to be taken, though it has been over a year since the violence.²⁵ In February of 2024 a woman was nearly lynched by a mob for merely wearing clothing with text written in

Arabic.²⁶ 2024 has seen other blasphemy cases be reported in Chichawatni and Lahore, both cities in the Punjab Province. In Chichawatni, a man was accused of blasphemy after a personal conflict escalated, while in Lahore a teenage boy faced blasphemy charges for allegedly sharing objectionable content on social media. Both incidents have led to public outcry and swift legal actions, underscoring the sensitive nature of these accusations and their potential to quickly escalate into greater communal tensions.²⁷

The current draconian legislation relating to blasphemy in Pakistan has created a chilling effect on free speech, suppressing political activism and critical journalism. Journalists face immense danger when discussing religious issues, as even perceived blasphemy or criticism of religious groups can lead to death threats or worse. A prominent Pakistani journalist, Raza Rumi, narrowly escaped an assassination attempt after criticizing the misuse of blasphemy laws in Pakistan.²⁸ The Clooney Foundation for Justice (CFJ) observed 24 blasphemy cases over a span of 6 months in 2022 and 252 resultant hearings. The CFJ report also highlights the devastating effects of blasphemy legislation in Pakistan, which is often abused due to vague definitions and a lack of procedural safeguards. According to the report, many blasphemy cases are filed without substantial evidence, resulting in prolonged detentions and significant delays in trials.²⁹

In this social and historical context, studies by institutions such as the Oxford Internet Institute and EU DisinfoLab show how social media algorithms, which favour emotionally charged and sensational content, often provide a breeding ground for hate speech and extremist mobilisation. State and non-state actors have been found to exploit social media platforms to manipulate public sentiment, polarise communities, and organise violence. The intersection of blasphemy laws and social media has thus created a volatile environment where accusations can rapidly spiral into violence, with devastating consequences for the accused and the broader society.

Party Y & Their Online Political Attack on Women Journalist Z

On May 25, 2024, Nazir Masih, an elderly Christian resident of Sargodha, died at the hands of mob violence, as a result of blasphemy allegations. Nazir Masih and his son were accused of scattering Quranic pages in front of their shoe factory in Mujahid Colony.³⁰ The incident highlighted deep-seated, historical problems of religious intolerance, mob or vigilante justice and the lack of legal protections in place for minority communities in Pakistan, as well as the inability of the police to proactively and successfully intervene on a number of occasions. According to the Human Rights Commission of Pakistan (HRCP)'s fact-finding mission, the Sargodha incident was not an isolated or spontaneous event, but a targeted attack against the Christian community.³¹ The scale of the incident, with police identifying 44 individuals and booking 500 in connection with the attack, further highlights the alarming prevalence of mob violence fueled by religious discrimination.³²

A well-respected Pakistani woman journalist, whom we shall refer to from this point onwards as "Z" (minus quotations), tweeted in response to the incident. Her

tweeting in turn led to an avalanche of online backlash and abuse on the social media platform X (formerly known as Twitter), particularly from a religious political far right party we will refer to from this point on as “Party Y”. According to Z’s tweet, *“One of the main challenges faced by Pakistan currently is radicalism at the hands of a far-right religious extremist party. Party Y, which is state-backed, has arguably transcended the state’s control capabilities, and continues to expand in influence and numbers. The party targets the young in particular – unemployed boys whose focus is then dangerously diverted towards the affairs of religious minorities.”*³³ This tweet by the woman journalist resulted in an alarming reaction from the supporters of Party Y, who demanded either an apology or proof of her accusatory claims against the party.

In response to Z’s critical commentary, Party Y’s supporters launched a coordinated online campaign aimed at discrediting and silencing her. The attack manifested in a number of ways. As mentioned, Party Y’s supporters flooded Z’s social media accounts with demands for an apology or concrete evidence supporting her claims: this tactic sought to undermine her credibility as a journalist by framing her statements as unfounded. The messages directed at Z contained menacing undertones, with some users employing aggressive and intimidating language. There were also tweets directing death threats at her. Moreover

the online abuse was heavily gendered, going beyond professional criticism. Z faced misogynistic remarks, derogatory comments about her appearance, and sexist insults. Party Y’s supporters adeptly used social media features like hashtags, mentions, and coordinated timing to amplify their messages. The strategic and coordinated use of the platform’s algorithms increased the visibility of their attack, potentially inciting further harassment from a broader audience.

This coordinated campaign against Z mirrors the dark reality of Pakistan’s social media landscape and lack of accountability of social media platforms, where these forms of online harm are rampant.

The rising trend in targeted hate speech against journalists in Pakistan across social media platforms has been particularly significant since the Sargodha incident. DRF has observed an increase in mobilisation via Twitter spaces, Tiktok accounts, violent and/or graphic videos on Facebook and YouTube declaring religious war critics (perceived and actual), particularly propagating hate speech and coordinated campaigns against vulnerable professions and religious minorities.

Theoretical Framework

For this particular case study, DRF employed two theoretical frameworks to provide a foundation for understanding the complex dynamics of online gendered hate speech in Pakistan.

Kimberlé Crenshaw's theory of intersectionality

Kimberlé Crenshaw's intersectionality theory is crucial in understanding how women in Pakistan experience multiple layers of oppression. The intersection of gender and profession (being a journalist) significantly amplifies the risks women face in public discourse. In Pakistan, a deeply patriarchal society, this effect is compounded by the religious conservatism that frequently targets women journalists who critique socio-political norms. DRF's study on *Online Violence Against Women in Pakistan* reveals that women journalists and activists are particularly vulnerable to online harassment, including death threats, character assassination, and accusations of blasphemy, which further heighten the risks due to Pakistan's strict blasphemy laws.^{34 35} This illustrates how gender and religious identity interact to expose

women to more severe consequences, both socially and legally.³⁶

Laura Mulvey's concept of the "male gaze"

Laura Mulvey's concept of the "male gaze" sheds light on the patriarchal control over public discourse often results in women being objectified and scrutinized more harshly as compared to their male counterparts, particularly when they challenge dominant norms. Bytes for All's *I Don't Forward Hate* campaign highlights the increasing trend of hate speech targeting women and minorities in Pakistan's digital spaces. The campaign documents incidents where women, especially those in public-facing professions, face misogynistic hate, and religious intolerance, illustrating how the male gaze continues to shape their experiences in the public domain.³⁷ Women journalists such as Z not only face criticism for their professional opinions but also for violating the societal expectations imposed on their gender.

Methodology

Owing to privacy and security concerns pertaining to the sensitivity of the case, identities of the entities involved have been anonymised. Therefore, throughout this report, the woman journalist would be referred to as Journalist Z and the far right religious party would be labelled as Party Y, as we have noted earlier in the report and in the disclaimer. The study focuses on a quantitative investigation of hate speech based on the Sargodha incident and the response to the remarks made by Z. Tweets on X were examined to offer insights into the sentiments and attitudes expressed online within a sample of the population. By employing mathematical and statistical analytical techniques, the study examined the prevalence and distribution of themes of hate speech, and identified any patterns and trends in hate speech discourse.

The tweets collected – a total of 565 – showcase the level of discourse and polarization evident on this platform, serving as a salient feature for the analysis. Keywords associated with Journalist Z and Party Y around the timeframe of the Sargodha incident were identified and added to a curated list of keywords. For a comparative gendered analysis of the case, tweets mentioning a male journalist within the same context were also compiled. The final list of keywords consisted of both English and Urdu words/phrases. In terms of data

collection, the study relied on collecting tweets consisting of the associated keywords. However, with paid and limited access to the data that can be collected using X's API key and the API key itself being behind a pay-wall of monthly subscriptions, alternative methods of data collection were explored.

As platform X has a gray policy concerning web scraping, and has put in place extensive measures to curb web scraping of tweets, the programming language Python was used to create custom tools that involved multiple steps to overcome each limitation. The most basic form of API access on X is US\$100 per month, and even at this cost one can only access data from the past week or seven days. Apart from having to pay for API access, the platform also has mechanisms in place to stop anyone from downloading the HTML version of the website. Once a method was developed to bypass this check, X placed another limitation to not include anything other than what is being displayed on the screen as it is. For example, if a user scrolls down the page, previous sections that were scrolled past are unloaded and taken out of the HTML file. In order to overcome this limitation, the Python tool zoomed out the page further than the screen's resolution, in order to load as many tweets as was possible at one time, onto a single screen. However, even with this structure, only limited data from the past 7 days can be accessed. Such arduous limitations placed by the

platforms themselves pose significant challenges for researchers and academics trying to study social media platforms and the discourses that take place on these platforms, across numerous academic and research disciplines.

Once associated keywords were assimilated, the HTML files of the search result pages for each keyword were separately downloaded. A Python script was then developed to parse the HTML files and extract relevant data. After successfully converting the HTML file into readable text format using the 'beautifulsoup' Python library, the data was cleaned and sorted, using the 'pandas' Python library, to remove any duplicates or manage missing values from the dataset.^{38 39} The aforementioned total of 565 tweets were collected into a CSV file. It is important to note here that if two users posted the same text tweet, it was not considered a duplicate tweet; rather, they were treated as two separate entities. The final, cleaned and processed dataset consisted of 500 unique posts. The data was then exported into Google Sheets, where it was annotated based on predefined variables. HTML files were used as a reference during annotations to understand any image or video based-contexts for the tweet text. This holistic approach ensured that the annotations contained full contextual understanding and were not judged solely on the basis of the text extracted from the tweets. After annotations, the final dataset was reviewed twice to remove any individual

biases.

However, the relatively small sample size of the dataset means that the findings may not fully capture the diversity of perspectives and opinions within the broader population. This limitation restricts the generalizability of the results beyond the specific event and context examined in the study, as hate speech dynamics can vary significantly depending on specific circumstances and sociopolitical factors. Additionally, as the data collection and annotation processes were conducted manually to ensure accuracy, reliability, and relevance to this case study, it is possible that some human subjectivity may have been introduced, which is difficult to fully eliminate. While this may introduce random biases in the analysis, peer review was done to minimize such biases and maintain the integrity of the data. By using specific keywords and hashtags relating to the case study, tweets were extrapolated and analyzed to develop contextual understanding in the form of the following variables in the tables below.

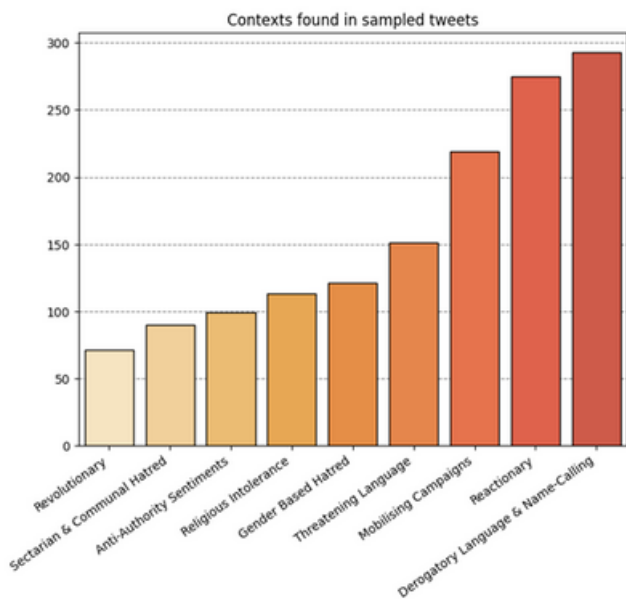
Note: While "Religious Intolerance", "Sectarian and Communal Hatred", "Derogatory Language and Name Calling", and "Threatening Language" may appear to overlap, and indeed do so to an extent, there are also points of distinction. "Sectarian and Communal Hatred" may include language that is discriminatory in terms of religious belief, but can also include racial, social, or regional grounds for discrimination, which "Religious Intolerance", as defined below, may not. Similarly, "Derogatory Language and Name Calling", and "Threatening Language" share common traits, but the former does not necessarily include inciting and promising to act upon threats to an individual or community.

Variable	Description
Religious Intolerance	Calling out or discriminating against individuals and communities on the basis of their religion or religious sympathies.
Sectarian and Communal Hatred	Discriminating, spreading hatred or inciting violence against marginalized groups on the basis of region, ethnicity, religion, race, community, social affiliations, etc.
Threatening Language	Using language that intimidates, threatens, or incites harm against individuals or communities based on their religious, social, political, ideological, racial, economic, and/or cultural differences. This includes direct threats, ultimatums, and warnings aimed at creating fear or compliance.
Mobilizing Campaigns	Organizing or encouraging collective action, protests, or movements against particular individuals or communities due to differences in social, religious, political, ideological, racial, economic, and/or cultural perspectives. This includes rallying followers to participate in activities that may lead to social unrest or violence.
Derogatory Language and Name Calling	Passing derogatory, condescending, insulting, passive aggressive, and/or misogynistic remarks against individuals/communities on the basis of religious, social, political, ideological, racial, economic, and/or cultural differences.
Anti-Authority Sentiments	Sharing comments, opinions or passing condescending, derogatory, discriminatory or hateful remarks against the state and/or authority figures, including but not limited to policymakers, bureaucrats, office holders, government officials, military and/or anyone belonging to state institutions.

Reactionary	Expressing biased, prejudiced, or conservative opinions rooted in a desire to maintain traditional or existing social, political, or ideological structures. This includes supporting established norms and opposing progressive changes or movements, often in defense of a particular religion, sect, political party or candidate.
Revolutionary	Expressing support for significant changes or reforms to existing social, political, or ideological structures. This includes advocating for progressive ideas, policies, or movements that seek to improve or transform current systems in favour of new approaches, often aligned with secular views and/or a particular political party or candidate.
Gender Based Hatred	Passing derogatory, condescending, and/or misogynistic remarks against individuals based on their gender with the potential intention of inciting hatred and/or violence, taking into account their allied interests towards a particular gender or gendered community.

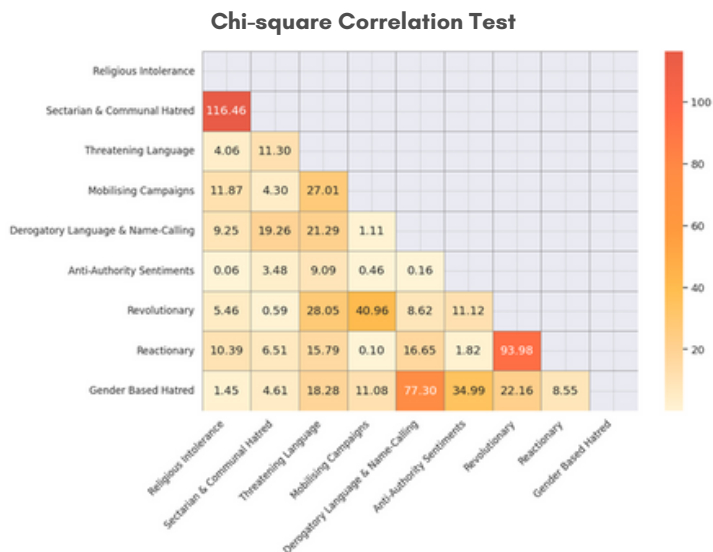
In the context of this study, the variables are defined with particular emphasis on the specific nuances relevant to our research focus. While these definitions are tailored to suit the objectives of this study, they remain consistent with the basic and rhetorically accepted definitions recognized worldwide, as found in authoritative sources such as the Oxford English Dictionary, Merriam-Webster, and Collins English Dictionary. This ensures both the relevance and validity of the variables in addressing the research questions.

Findings & Analysis



Exploratory data analysis

A total of 565 tweets were collected based on the defined methodology. After processing and cleaning the dataset, the final sample consisted of 500 unique posts. The exploratory data analysis of the documented tweets reveals that posts containing “Revolutionary” language were 73. “Sectarian and Communal Hatred” was found in 92 tweets whereas “Anti-Authority Sentiments” were in 102 tweets. 117 tweets comprised “Religious Intolerance” followed by “Gender Based Hatred” in 124 tweets. “Threatening Language” was found in 153 tweets and 223 tweets contained content related to “Mobilising Campaigns”. “Reactionary” content was found in 277 tweets whereas “Derogatory Language & Name-Calling” was the highest at 296 tweets. It is important to note that multiple variables had the propensity to exist within the same tweet as shown by the bar plot.



The Chi-square correlation test was conducted to detect any noticeable correlation between the pairings of variables. The test concludes with a few interesting connections, such as high correlations between the following pairings: ‘Sectarian & Communal Hatred’ and ‘Religious Intolerance’, ‘Gender Based Hatred’ and ‘Derogatory Language & Name-Calling’, ‘Revolutionary’ and ‘Mobilising Campaigns’, ‘Gender Based Hatred’ and ‘Anti-Authority Sentiments’. We will delve further into the directionality of the correlations among these selected groups by using contingency tables.

It is imperative to note that we have excluded the high correlation between the variables ‘Revolutionary’ and ‘Reactionary’ as intuitively we would expect these variables to have a high negative correlation, based on the antagonistic nature of the definitions of our variables. If a tweet is found to have a significant political sentiment, it can either be ‘Revolutionary’ or ‘Reactionary’ but not both.

Contingency Table for 'Sectarian & Communal Hatred' and 'Religious Intolerance'

	Sectarian & Communal Hatred Present	Sectarian & Communal Hatred Absent
Religious Intolerance Present	59	54
Religious Intolerance Absent	29	355

This contingency table shows the frequency counts of cases where each combination of 'Sectarian & Communal Hatred' and 'Religious Intolerance' is present or absent. From the correlation heat map, we see that the Chi-square statistic value comes out at 113.15, which is the highest in the entire table. The p-value comes out to $3.761e-27$, which is considerably lower than the threshold value of 0.05 which indicates that the observed association between 'Sectarian & Communal Hatred' and 'Religious Intolerance' is highly statistically significant. This suggests that the relationship observed in the sample data is unlikely to have occurred by chance. The Phi coefficient comes out to be 0.484 which is between 0 and 1 thus indicating a positive association.

Therefore, the contingency table leads us to the conclusion that when 'Sectarian & Communal Hatred' is present, there is a higher likelihood of 'Religious Intolerance' being present, and vice versa.

Contingency Table for 'Gender Based Hatred' and 'Derogatory Language & Name-Calling'

	Gender Based Hatred Present	Gender Based Hatred Absent
Derogatory Language & Name-Calling Present	113	179
Derogatory Language & Name-Calling Absent	8	197

This contingency table shows the frequency counts of cases where each combination of 'Gender Based Hatred' and 'Derogatory Language & Name-Calling' is present or absent. From the correlation heat map we see that the Chi-square statistic value comes out at 78.17, which is the third highest in the table. The p-value comes out to $1.471e-18$ which is considerably lower than the threshold

value of 0.05 which indicates that the observed association between 'Gender Based Hatred' and 'Derogatory Language & Name-Calling' is highly statistically significant. This suggests that the relationship observed in the sample data is unlikely to have occurred by chance. The Phi coefficient comes out to be 0.4 which is between 0 and 1 thus indicating a positive association.

As a result, the contingency table leads us to the conclusion that when 'Gender Based Hatred' is present, there is a higher likelihood of 'Derogatory Language & Name-Calling' being present, and vice versa.

Contingency Table for 'Revolutionary' and 'Mobilising Campaigns'

	Revolutionary Present	Revolutionary Absent
Mobilising Campaigns Present	6	213
Mobilising Campaigns Absent	65	213

This contingency table shows the frequency counts of cases where each combination of 'Revolutionary' and 'Mobilising Campaigns' is present or absent. From the correlation heat map we see that the Chi-square statistic value comes out at 40.15. The p-value comes out to 1.557e-10 which is considerably lower than the threshold value of 0.05 which indicates that the observed association between 'Revolutionary' and 'Mobilising Campaigns' is highly statistically significant. This suggests that the relationship observed in the sample data is unlikely to have occurred by chance. The Phi coefficient comes out to be -0.137 which is between -1 and 0 thus indicating a negative association.

Therefore, the contingency table leads us to the conclusion that when the variable 'Revolutionary' is present, there is a higher likelihood of 'Mobilising Campaigns' being absent, and vice versa.

Contingency Table for 'Gender Based Hatred' and 'Anti-Authority Sentiments'

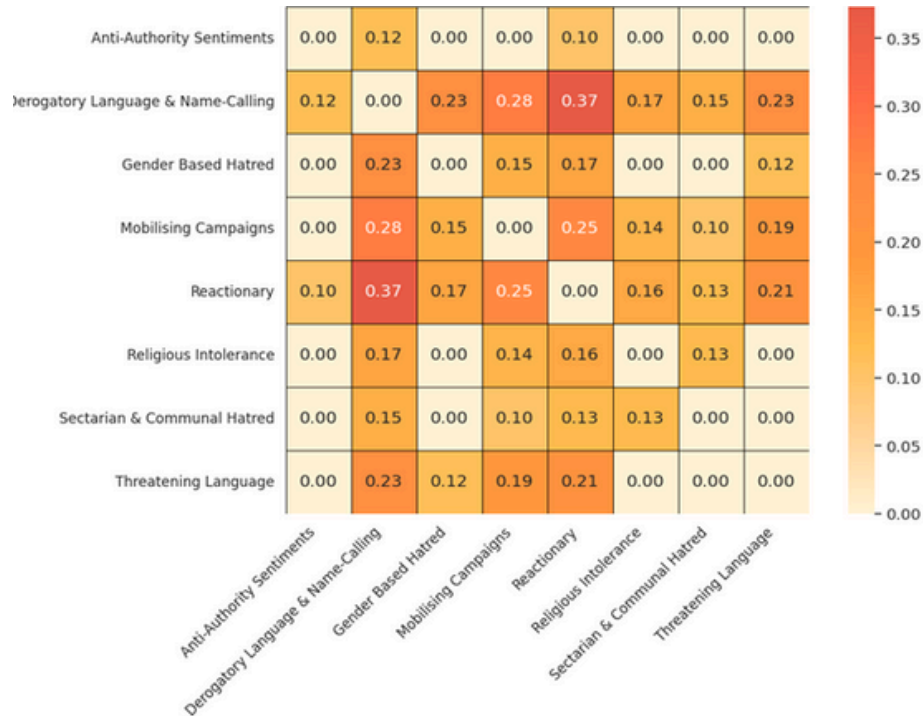
	Gender Based Hatred Present	Gender Based Hatred Absent
Anti-Authority Sentiments Present	1	98
Anti-Authority Sentiments Absent	120	278

This contingency table shows the frequency counts of cases where each combination of 'Gender Based Hatred' and 'Anti-Authority Sentiments' is present or absent. From the correlation heat map we see that the Chi-square statistic value comes out at 34.51. The p-value comes out to 3.321e-09 which is considerably lower than the threshold value of 0.05 which indicates that the observed association between 'Gender Based Hatred' and 'Anti-Authority Sentiments' is highly statistically significant. This suggests that the relationship observed in the sample data is unlikely to have occurred by chance. The Phi coefficient comes out to be -0.268 which is between

-1 and 0 thus indicating a negative association.

Thus, the contingency table leads us to the conclusion that when there is 'Gender Based Hatred' present, there is a higher likelihood of 'Anti-Authority Sentiments' being absent, and vice versa.

Pairwise Support Analysis Heatmap



The *a priori* analysis is conducted to further explore and quantify the specific patterns or rules governing the associations between sets of variables in our data.

confidence of 0.6 which is done to ensure that only strong and reliable associations are considered.

The support heatmap showcases the values of the pairwise support rule. From the map, 'Derogatory Language & Name-Calling' shows relatively high support with multiple pairings such as; 'Gender Based Hatred', 'Mobilising Campaigns' and 'Reactionary'. We also notice 'Reactionary' and 'Mobilising Campaigns' as having high support. Regarding 'Gender Based Hatred' we notice some existing support with 'Threatening Language'.

Additionally, we constructed a table showing the different association rules with their respective values. All significant associations displayed have a minimum

Note: For the following table: SCH = Sectarian & Communal Hatred, RI = Religious Intolerance, TL = Threatening Language, MC = Mobilising Campaigns, DLNC = Derogatory Language & Name-Calling, GBH = Gender Based Hatred

Antecedents	Consequents	Support	Confidence	Lift	Leverage	Conviction	Zhangs_metric
(SCH)	(RI)	0.126	0.685	2.932	0.083	2.432	0.807
(RI)	(DLNC)	0.168	0.718	1.215	0.030	1.451	0.231
(RI)	(Reactionary)	0.160	0.684	1.237	0.031	1.414	0.250
(SCH)	(DLNC)	0.148	0.804	1.361	0.039	2.091	0.325
(SCH)	(Reactionary)	0.126	0.685	1.239	0.024	1.418	0.236
(TL)	(MC)	0.192	0.627	1.410	0.056	1.489	0.418
(TL)	(DLNC)	0.230	0.752	1.272	0.049	1.647	0.308
(TL)	(Reactionary)	0.212	0.693	1.253	0.043	1.455	0.291
(MC)	(DLNC)	0.275	0.619	1.047	0.012	1.074	0.082
(DLNC)	(Reactionary)	0.373	0.632	1.143	0.047	1.214	0.305
(Reactionary)	(DLNC)	0.373	0.675	1.143	0.047	1.259	0.279
(GBH)	(DLNC)	0.232	0.935	1.583	0.085	6.342	0.490
(GBH)	(Reactionary)	0.168	0.677	1.225	0.031	1.386	0.244
(SCH, DLNC)	(RI)	0.102	0.689	2.951	0.067	2.466	0.776
(SCH, RI)	(DLNC)	0.102	0.810	1.370	0.028	2.148	0.309
(DLNC, RI)	(SCH)	0.102	0.607	3.306	0.071	2.078	0.838

(DLNC, RI)	(MC)	0.108	0.643	1.444	0.033	1.554	0.370
(RI, MC)	(DLNC)	0.108	0.771	1.306	0.025	1.790	0.272
(RI, MC)	(Reactionary)	0.106	0.757	1.369	0.029	1.841	0.314
(RI, Reactionary)	(MC)	0.106	0.663	1.488	0.035	1.644	0.390
(DLNC, RI)	(Reactionary)	0.118	0.702	1.270	0.025	1.502	0.256
(RI, Reactionary)	(DLNC)	0.118	0.738	1.248	0.023	1.559	0.237
(SCH, DLNC)	(Reactionary)	0.104	0.703	1.271	0.022	1.504	0.250
(SCH, Reactionary)	(DLNC)	0.104	0.825	1.397	0.029	2.343	0.325
(DLNC, TL)	(MC)	0.144	0.626	1.407	0.042	1.484	0.375
(TL, MC)	(DLNC)	0.144	0.750	1.269	0.031	1.637	0.263
(DLNC, TL)	(Reactionary)	0.166	0.722	1.305	0.039	1.607	0.304
(TL, Reactionary)	(DLNC)	0.166	0.783	1.325	0.041	1.886	0.311
(GBH, TL)	(DLNC)	0.114	0.983	1.663	0.045	23.733	0.451
(DLNC, MC)	(Reactionary)	0.170	0.616	1.114	0.017	1.164	0.141
(MC, Reactionary)	(DLNC)	0.170	0.675	1.142	0.021	1.257	0.166
(GBH, MC)	(DLNC)	0.134	0.918	1.553	0.048	4.978	0.417

(GBH, DLNC)	(Reactionary)	0.160	0.690	1.247	0.032	1.441	0.258
(GBH, Reactionary)	(DLNC)	0.160	0.952	1.612	0.061	8.593	0.456
(GBH)	(DLNC, Reactionary)	0.160	0.645	1.728	0.067	1.766	0.560

Based on the table, there are a few associations that we will highlight:

(1) Gender Based Hatred (GBH) → Derogatory Language & Name-Calling (DLNC)

This association has a confidence level of 0.935, meaning *when instances of Gender Based Hatred (GBH) are observed, there is a high likelihood (93.5%) that these instances will also involve Derogatory Language & Name-Calling (DLNC)*. A lift value of 1.583 indicates that the occurrence of DLNC is 1.583 times more likely when GBH is present compared to its expected occurrence by chance alone. A leverage value of 0.085 indicates that GBH and DLNC occur together more frequently than expected, further supporting the association between these behaviors. A high conviction value of 6.342 suggests that the presence of GBH strongly determines the presence of DLNC. Finally, Zhang's metric value of 0.490 indicates a moderate level of association between GBH and DLNC.

(2) Derogatory Language & Name-Calling (DLNC) and Religious Intolerance (RI) → Sectarian & Communal Hatred (SCH)

This association has a confidence level of 0.607, meaning *when instances of DLNC (Derogatory Language & Name-Calling) and RI (Religious Intolerance) are observed, there is a significant likelihood (60.7%) that these instances will also involve SCH (Sectarian & Communal Hatred)*. A lift value of 3.306 indicates that the occurrence of SCH is 3.306 times more likely when both DLNC and RI are present compared to its expected occurrence by chance alone. A leverage value of 0.071 indicates that DLNC, RI, and SCH occur together more frequently than expected, further supporting the association between these behaviors. A conviction value of 2.078 suggests that the presence of both DLNC and RI increases the likelihood of SCH occurring. Finally, Zhang's metric value of 0.838 indicates a strong level of association between DLNC, RI, and SCH.

(3) Gender Based Hatred (GBH) and Threatening Language (TL) → Derogatory Language & Name-Calling (DLNC)

This association has a confidence level of 0.983, meaning **when instances of GBH (Gender Based Hatred) and TL (Threatening Language) are observed, there is a very high likelihood (98.3%) that these instances will also involve DLNC (Derogatory Language & Name-Calling)**. A lift value of 1.663 indicates that the occurrence of DLNC is 1.663 times more likely when both GBH and TL are present compared to its expected occurrence by chance alone. A leverage value of 0.045 indicates that GBH, TL, and DLNC occur together more frequently than expected, further supporting the association between these behaviors. A high conviction value of 23.733 suggests that the presence of both GBH and TL strongly determines the presence of DLNC. Finally, Zhang's metric value of 0.451 indicates a moderate level of association between GBH, TL, and DLNC.

(4) Gender Based Hatred (GBH) and Reactionary → Derogatory Language & Name-Calling (DLNC)

This association has a confidence level of 0.952, meaning **when instances of GBH (Gender Based Hatred) and Reactionary behavior are observed, there is a very high likelihood (95.2%) that these instances will also involve DLNC (Derogatory Language & Name-Calling)**. A lift value of 1.612 indicates that

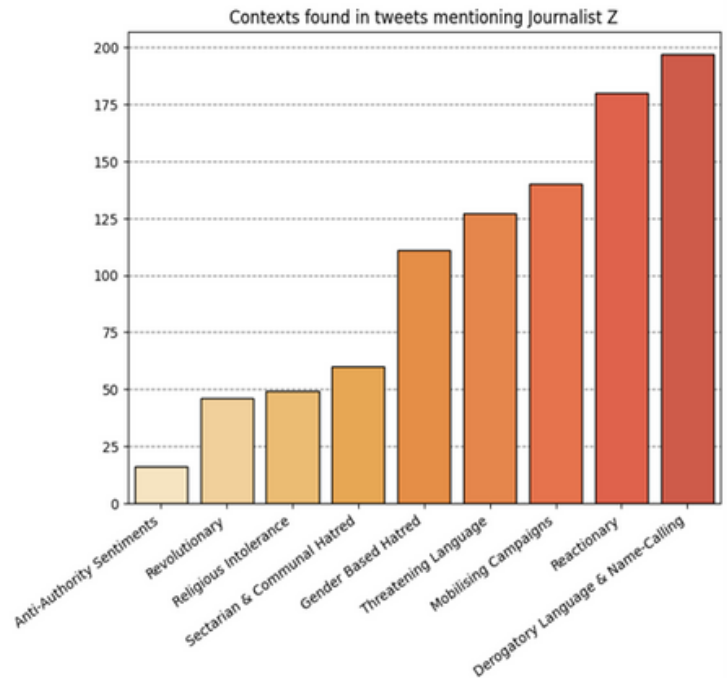
the occurrence of DLNC is 1.612 times more likely when both GBH and Reactionary behavior are present compared to its expected occurrence by chance alone. A leverage value of 0.061 indicates that GBH, Reactionary behavior, and DLNC occur together more frequently than expected, further supporting the association between these behaviors. A high conviction value of 8.593 suggests that the presence of both GBH and Reactionary behavior strongly determines the presence of DLNC. Finally, Zhang's metric value of 0.456 indicates a moderate level of association between GBH, Reactionary behavior, and DLNC.

Based on these analyses, several conclusions can be drawn. Derogatory language is notably prevalent in our dataset, indicating a toxic digital environment that fosters excessive hostility and aggression. Such interactions are likely to cause some individuals to feel alienated and experience harm in various forms, potentially normalizing and perpetuating cycles of abuse and intolerance. Gender-based hatred is also widespread within the data, often in conjunction with other variables, which may exacerbate inequalities in the digital space and reinforce harmful stereotypes. These dynamics could lead to a lack of diversity and inclusivity in online conversations. Furthermore, the relatively high joint presence of reactionary content, tweets related to mobilizing campaigns, and derogatory language, including threats, suggests that hostile

behavior is often coordinated rather than isolated. This behavior is particularly prevalent among individuals with reactionary rather than revolutionary perspectives, amplifying the reach and impact of polarizing views across various online groups.

The use of derogatory language and threats underscores that these viewpoints are not only imposed on others but also conveyed in obscene ways, making such behavior a constant expectation in certain digital contexts. The presence of sectarian hatred and religious intolerance suggests deep-rooted religious and communal divisions. This interaction could also imply a cyclical reinforcement of prejudice, where intolerance breeds further sectarianism. Overall, the analysis indicates the significant erosion of online civility, the potential for negative impacts on mental health, reduced participation in discussions, the amplification of extreme and conflicting views, and a notable presence of distrust in authoritative figures and governing policies.

We also looked at tweets that make direct mention of Journalist Z and to examine how the amount of context present in those particular tweets differs from the data we had seen up to now.

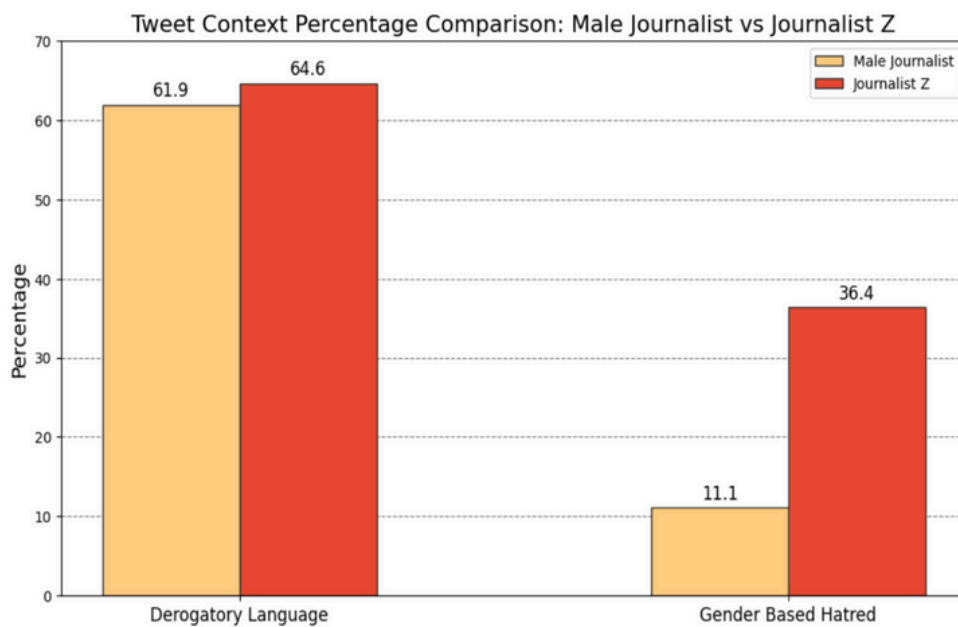


We identified approximately 309 tweets that mentioned Journalist Z, and among these, "Anti-Authority Sentiments" appeared in 16 tweets, "Revolutionary" themes in 46 tweets, and "Religious Intolerance" in 49 tweets. Furthermore, "Sectarian & Communal Hatred" was present in 60 tweets, "Gender-Based Hatred" in 111 tweets, and "Threatening Language" in 127 tweets. Additionally, "Mobilising Campaigns" were mentioned in 140 tweets, "Reactionary" themes in 180 tweets, and "Derogatory Language & Name-Calling" in 197 tweets. When comparing these figures to the percentages in the first barplot, tweets mentioning Z showed an increase of 11.72% in "Gender-Based Hatred," a 10.9%

increase in "Threatening Language," a 5.15% increase in "Derogatory Language & Name-Calling," and a 3.25% increase in "Reactionary" themes.

The data of Journalist Z was also compared to that of a male journalist, analyzing the frequency of 'Derogatory Language & Name-Calling' and 'Gender Based Hatred' contexts mentioned in tweets, which was then presented in a bar plot.

suggests a particular type of engagement context concerning women journalists, although it should be noted that this observation is limited to this specific case. It is, however, important to consider that the nature of the abuse that Z faced, including misogynistic slurs and threatening language, highlights how women journalists are uniquely targeted. The comparative data results can be viewed through Kimberlé Crenshaw's



This exploratory data analysis highlights the relatively heightened gender-based discrimination experienced, as well as a considerable increase in online threats. This may be a compounded effect related to the gender of the journalist or due to the individual's occupation and influence in specific digital circles, although the gender comparative barplot shows a considerably noticeable difference in the gender based hatred noticed by Z. The increased presence of derogatory language and reactionary content

theory of intersectionality, as they highlight an important framework to understand the exacerbated discrimination that journalist Z has been subjected to. The theory suggests that various types of inequalities and marginalizations in compounded form lead to barriers not understandable through a single, conventional lens.⁴⁰ Z's gender and profession made her vulnerable to multifaceted oppression. The backlash she faced showcased systemic gender biases aimed at silencing

women and challenging dominant groups in society. Studies indicate that women journalists are more likely to experience online harassment than their male counterparts, with threats often focusing on their gender and aiming to discredit their professional credibility and reinforce patriarchal norms.^{41 42}

We also see how other theoretical frameworks, such as Laura Mulvey's concept of the 'male gaze' showcase women being objectified and subjected to a patriarchal perspective in media representations.⁴³ This framework helps us understand the heightened presence of gender-based hatred and derogatory language directed towards women, especially female journalists like Journalist Z. The significant increases in such hostile mentions highlight the entrenched patriarchal norms that seek to undermine women's credibility and presence in public discourse. The 'male gaze' transforms the digital space into a battleground where women's voices are continually marginalized and attacked.

Similarly, Spivak's work on the subaltern is significant as it emphasizes the importance of bringing the voices of the marginalized, especially women in Pakistan to the forefront. Additionally, the blasphemy laws are a postcolonial product, the violence that manifests through them affecting the most marginalized in society, reflecting Spivak's work which is influenced by the notion that European colonialism still affects the

social, economic, and political life of those in postcolonial nations.⁴⁴ Moreover, Critical Discourse Analysis (CDA) elucidates the power dynamics and ideologies embedded within our data. The analysis of tweets shows how derogatory and intolerant language functions to reinforce existing power structures and perpetuate societal divides. For example, the significant correlation between sectarian/communal hatred and religious intolerance indicates how discourse can perpetuate communal divisions and maintain hegemonic ideologies. CDA reveals that language is not neutral but a tool of power that can shape and reinforce social inequalities.⁴⁵

In addition to the above findings, we observed events being conducted in online spaces, particularly X by Party Y with the intent to emanate their reactionary ideologies. The hosted events included online conversations on issues promoting intolerance against religious minorities. As a consequence, two men from a religious minority community were shot dead by a nineteen-year-old student. After the accused was arrested, the District Police Officer (DPO) stated that the student confessed to the murders.⁴⁶ The accused also confessed that the killings were a result of him watching religious content against the minority community on social media platforms. He added that he wished to kill every person belonging to the particular religious minority community.⁴⁷ According to the official press release put out by the

Punjab Police, the community spokesperson stated, *“a hate campaign had been launched for the last few weeks ahead of Eid ul Azha in a bid to stop their community from observing the occasion... hate videos were viral on social media against their community over observing Eid ul Azha.”*⁴⁸

In another incident, a tourist accused of blasphemy in Swat, Khyber Pakhtunkhwa, was burned alive in a violent mob attack.⁴⁹ Online hate speech, derogatory remarks, and religious and communal hatred spread by reactionary elements hold significant influence over the public, leading to violent attacks. Due to poor content moderation policies, such content continues to be widespread on platforms, enabling toxic online spaces and resulting in considerable offline harm to marginalized communities such as women and religious minorities. The stronghold of such political groups and their political and religious ideologies leading to such attacks is a grave concern, as they show the demonstrated ability of the violence they are capable of spreading. The serious consequences are not only targeted at a single group in society but have the potential to harm any dissenting voice including both psychological and physical harms.

In order to further understand the psychological impact and the potential of such threatening hate campaigns to translate to physical harm, we reached out to journalist Z. While expressing grave

concern about the situation, she added that the experience was the first of its kind for her, as the online harassment she had been subjected to previously did not pose a threat to her life. The following transcript is of her expressions and emotions regarding this event:

“The incident with Party Y has been an intense chapter in my professional career (as a journalist). Since I am opinionated about politics on social media, we usually have to face significant online harassment and trolls on the platforms however the intensity of the interactions and backlash is not as much as this incident as it was being linked to a religious occurrence and this is a radical religious group. What was different this time around was that the backlash included several threats.

“There was a tweet that... stated that we can easily transform from a flower to a sword. I received several tweets and inbox messages of this nature as well...”

Although there were a lot of gender based threats, the majority of them were from a religious point of view. For example, there was a tweet that I also responded to. It stated that ‘we can easily transform from a flower to a sword’. I received several tweets and inbox messages of this nature as well however I did not highlight them as much. There was a tweet that gave the example of how ‘someone in Europe, a Party Y worker,

"Although I was horrified due to this incident, I have still been very consistent in writing about it... [and] raising my voice (against the atrocities...)"

killed someone, warning me that we do not even spare anyone in a foreign country so you should watch out for yourself.' So this was a first of its kind experience for me because I have always been very opinionated but I have never called out anyone directly. So in my criticism, not only did I mention Party Y but also the organization's founder [name redacted for privacy and safety] drawing a comparison between him and his workers/supporters.

Although I was horrified due to this incident, I have still been very consistent in writing about it. I am still raising my voice (about the atrocities at the groups' hands against minority communities). This was different compared to the online harassment we face by other political parties. Again, there were also instances like for example if they found a photo of

"There was a tweet that gave the example of how someone in Europe, a 'Party Y' worker, killed someone, warning me that we do not even spare anyone in a foreign country so you should watch out for yourself."

me from Instagram in which I was wearing a skirt, they posted that photo and passed gender based remarks on them, the kind that Pakistani men usually say. I think that its intensity was a lot as it also makes you fear for your own life.

I have some friends who supported me and till today, the posters that Party Y has created, they have also added my friends' photos in them. They are calling us Qadiyani,⁵⁰ atheists and other such terms. There are many Youtube shows (that) they did, there were Facebook live programs that the Party Y guys did

"...its intensity was a lot as it also makes you fear for your own life."

against me where my images were used. And yes, the women who appear on media (platforms), their characters were assassinated. So yes, it was a very intense experience for me but I also learned a few things that you should talk about issues but also talk between the lines. Even if you want to take a position (against issues) openly, you should do it a certain way and see what kind of people you should not respond to. So I also got the chance to learn this through the incident but I have been very vocal about the religious radicalism in Pakistan and I will still keep speaking about it."

Platform Policies: Addressing Gendered Hate Speech and Violent Content

According to X's documented community guidelines, the hate campaign against journalist Z potentially violates four content moderation policies.

Abuse and Harassment	<ul style="list-style-type: none">• You may not target others with abuse or harassment, or encourage other people to do so.• Some posts may appear to be harmful when viewed in isolation, but may not be when viewed in the context of a larger conversation.• We prohibit the malicious, unreciprocated targeting (such as mentioning or tagging) of individual(s), particularly when shared to humiliate or degrade someone. This can mean:<ul style="list-style-type: none">◦ Sharing multiple Posts, over a short period of time, or continuously posting replies with malicious content, to target an individual. This includes accounts dedicated to harassing an individual or multiple individuals.◦ Mentioning or tagging users with malicious content.• Anyone can report violations of this policy using our dedicated reporting flow. However, we sometimes need to hear directly from the person being targeted to ensure that we have the information needed prior to taking any enforcement action.• To facilitate healthy dialogue on the platform, and empower individuals to express diverse opinions and beliefs, we prohibit behavior and content that harasses, shames, or degrades others. In addition to posing risks to people's safety, these types of behavior may also lead to physical and emotional hardship for those affected.
-----------------------------	---

Hateful Conduct

- You may not directly attack other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.
- We will review and take action against reports of accounts targeting an individual or group of people with any of the following behavior, whether within Posts or Direct Messages.
 - Hateful references
 - Incitement
 - Slurs and Tropes
 - Dehumanization
 - Hateful imagery
 - Hateful profile
- Under this policy, we take action against behavior that targets individuals or an entire protected category with hateful conduct, as described above. Targeting can happen in a number of ways, for example, mentions, including a photo of an individual, referring to someone by their full name, etc.
- When determining the penalty for violating this policy, we consider a number of factors including, but not limited to the severity of the violation and an individual's previous record of rule violations.

<p>Violent Speech Policy</p>	<ul style="list-style-type: none"> • You may share graphic media if it is properly labeled, not prominently displayed and is not excessively gory or depicting sexual violence, but explicitly threatening, inciting, glorifying, or expressing desire for violence is not allowed. • Violent Content is any content containing Violent Speech or Violent Media, as defined below: <ul style="list-style-type: none"> ◦ Violent Speech: Content that threatens, incites, glorifies, or expresses desire for violence or harm. ◦ Violent Media: Visual material depicting graphic, violent or excessively gory content including sexual violence. • Anyone can report unmarked content or other violations using our dedicated in-app reporting flow.
<p>Violent and Hateful Entities Policy</p>	<ul style="list-style-type: none"> • There is no place on X for violent and hateful entities, including (but not limited to) terrorist organizations, violent extremist groups, perpetrators of violent attacks, or individuals who affiliate with and promote their illicit activities. • Violent entities are those that deliberately target humans or essential infrastructure with physical violence and/or violent rhetoric as a means to further their cause. These include, but are not limited to, terrorist organizations, violent extremist groups, and perpetrators of violent attacks. • Hateful entities are those that have systematically and intentionally promoted, supported and/or advocated for hateful conduct, which includes promoting violence or engaging in targeted harassment towards a protected category.

These four community guidelines discourage users from engaging in violent speech, abusing and/or harassing others, or spreading hate based on people's gender and gender identity, among other sociological categories. X also discourages violent and hateful entities' presence on the platform. However, despite these community guidelines being in place, Z has been continuously subjected to immense hatred and harassment on the platform by party Y's supporters and workers.

A question arises as to how users that are spreading hatred and threatening violence - including death threats - have been able to bypass X's content moderation policies. According to DRF's Helpline Report (2023), half of the cases escalated to X were left unresolved. The escalated cases were between January 2023 and May 2023. The report stated, "After the recent changes in X's (formerly Twitter) internal structure, the escalation channel completely disintegrated after May, although they previously exhibited an appreciable level of cooperation".⁵¹

Recommendations

Prior to Twitter's rebranding and restructuring, there were already concerns concerning threats to vulnerable communities on the platform. The dangers that Z, other women journalists, activists and other members of vulnerable communities highlight that rather than being tackled, bad actors have been enabled to exploit X and other platforms to amplify technology-facilitated gender-based violence (TFGBV), with seemingly no substantial penalties, such as blocks or bans. It is vital that X takes swift action against bad actors by strengthening and proactively enforcing their community guidelines and policies. To this end, we recommend the following steps to be taken:

01 Revive The Trusted Partner

Program: X needs to revive its trusted partner program in order to better moderate the content being shared on their website. With a lack of contextual and cultural understanding of social media platforms in terms of content moderation across diverse geographical regions, trusted partner programs have and continue to be effective measures to escalate cases often missed by the content moderation algorithms and platform's internal human reviewers. Thus, escalation channels need to be reinstated with civil society

organizations, particularly organizations working outside of the European Union where DSA promotes platform accountability.

02 Enhanced Transparency around Content Moderation:

There needs to be greater transparency regarding community guidelines and how content moderation is done on the platform. For example, according to X, what/who are the Dangerous organizations and Individuals? The list of organizations and individuals should be shared with the trusted partners in order to ensure a more streamlined and efficient reporting and escalation process.

03 Content Moderation Policies and

Methods: Our study observed that the tweets collected went against X's content moderation policies, with at least one and mostly more than one content policy being violated by the users during this organized campaign against Journalist Z. Party Y in question has continued to post content that surpasses the ambit of free speech and is promoting violent ideology, hate speech and using threatening language against an individual and communities with protected characteristics. Under such circumstances, X needs to provide better clarification on its content moderation policies and

limit the language of their policies to include varied regional contexts. For example, according to X's policies images of a person in a skirt may not directly violate the user privacy policy, however in a culturally conservative country like Pakistan, a mere picture in a skirt of a woman can be used to defame, discredit and at times be a threat to life for an individual. Therefore, the onus falls on X to ensure user safety and be vigilant in taking down non consensual use of an individual's private data in the public sphere.

04 Automated Content Moderation:

— According to X, the process of reporting, flagging and removing content is carried out through an automated system. However, there is no information provided on the automated system itself. If, like other social media platforms such as Meta, X uses natural language processing (NLP) within large language models (LLMs) that supposedly have the ability to remove content in all languages, the system still lacks robust mechanisms to identify and take down harmful content. X needs to a) be transparent in the automated system that it uses and b) include additional mechanisms to increase content moderation accuracy for languages other than English.

05 Beyond Existing Measures: X

— would have to develop better content moderation mechanisms to rely on refined AI technologies that detect and mitigate hate speech, including gender-based harassment, in real time. Such algorithms need regular training and updating to learn the subtleties of language and cultural contexts and be able to recognize noxious content properly. Next-generation systems use improved NLP techniques to understand context, sarcasm, and regional dialect much better, thus increasing the accuracy of harmful content detection.

06 Transparency around Content Moderation Algorithms:

— X should disclose the criteria and processes used by its algorithms to filter and prioritize content. Transparency is vital in many respects: to gain trust from users and stakeholders, to allow for audits by external bodies, and also to conduct oversight in an independent manner. Publicly available transparency reports should not only outline what content was removed and why but also demonstrate the effectiveness of moderation efforts. Such reports should present metrics about the false positive and false negative rates of AI systems used, as well as demographic breakdowns of affected users, to ensure the algorithms are not created at the

expense of specific demographics. The European Council insists on transparent, accountable—self and co-regulatory mechanisms in ensuring respect for human rights and the upholding of trust in society.⁵² For example, algorithmic accountability is articulated in the recommendation for meeting the requirements of transparent, explainable, and fair AI systems in the domains of content moderation. As reported, artificial intelligence could be effectively used to moderate content where human supervision would reduce hate speech, thus making the platform much safer and more trusted. Furthermore, transparency in algorithmic decision-making underpins the principles on which the EU's DSA is based, which seek more accountability within platforms' governance. Similarly, X should ensure that countries that are not part of the EU and the Global North are not left out from the conversations around platform accountability.

07 **Response and Resolution:**

— Effective reporting mechanisms enable users to respond to abuse effectively and afford platforms to assume that their response is prompt in a harassment case. Reported evidence shows that when reporting tools are satisfactory to use, they create trust from the users within the platform and enhance a safe environment online. X can also enable a team of moderators to be experts in catering to complaints of gender-based hate speech. This team should also pay attention to the cases of severe harassment in varied regional contexts and communicate updates on processing the reports on time to the users. The resolution process has to be transparent, giving clear indications to the user of any action taken and the decisions based on which such action has been concluded.

08 **Regular Impact Assessments and Independent Audits:**

— X should carry out large-scale impact assessments on how effective and efficient the content moderation policy has been in reducing cases of gendered hate speech. These could be received through user feedback and data analysis, as conducted by independent experts. Such tests should measure the effect moderation practices have on different demographic groups to ensure sound enforcement across

the board. There should be periodic third-party independent audits regarding how X's content moderation systems are carried out. These audits, the results of which will be publicly available, can inform how to frame policies better. The said audits shall result from implications surrounding algorithmic biases, the consistency of its application, and user satisfaction relating to the moderation process.

09 Independent Oversight Body:

Establish independent oversight bodies with representatives from CSOs, academics, and competent industry professionals for regular reviews of X's policies, with suggestions for changes. The boards should be accommodated with automatic, compulsory assessments connected to regular suggestions towards better ways to practice moderation. A clear advantage of CSO collaboration is that it enhances the effectiveness of platform policies and helps ensure policy content is sure to be responsive to the needs and rights of users. The principles of Participatory Governance emphasize the importance of stakeholder engagement in policy development and implementation. X should allow for the inclusion of various voices of CSOs to ensure that content moderation policies reflect an inclusive and culturally sensitive

input, addressing the uniqueness of the challenges that differences in these communities pose.

10 Facilitate Data Access for Independent Research:

Social media platforms like X should implement mechanisms for secure, anonymized access to platform data, specifically related to content moderation practices, hate speech incidents, and algorithmic decision-making. The current pay to access public data model of X, limits the already constricted resources of CSOs and NGOs, especially in developing countries. Therefore, the data should be made available to researchers, civil society organizations, and policymakers to promote transparency and enable independent evaluations of how hate speech, harassment, and platform policies affect vulnerable groups. By ensuring accessible data streams, platforms can foster collaborative efforts to refine content moderation approaches, support evidence-based policy development, and enhance accountability in managing online harm.

Bibliography

1. Asfandiyar (2023, June 21). *Pakistan's long history of Throttling Press Freedom*. The Diplomat. <https://thediplomat.com/2023/06/pakistan-long-history-of-throttling-press-freedom/>
2. G. Aceto, A. Botta, A. Pescapé, M. F. Awan, T. Ahmad and S. Qaisar, "Analyzing internet censorship in Pakistan," *2016 IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI)*, Bologna, Italy, 2016, pp. 1-6, doi: 10.1109/RTSI.2016.7740626.
3. Irfan, A. (2023, June 12). *Censoring the Media in Pakistan*. Instick. <https://inkstickmedia.com/censoring-the-media-in-pakistan/>
4. Pakistan Electronic Media Regulatory Authority <https://www.pemra.gov.pk>
5. *Country: Pakistan*. Reporters San Frontiers. <https://rsf.org/en/country/pakistan>.
6. Staff, C. (2024, October 10). *Pakistani authorities detain journalist after political reporting*. Committee to Protect Journalists. <https://cpj.org/2024/10/pakistani-authorities-detain-journalist-after-political-reporting/>
7. Stroud, S. (2022, June 27). *Censorship in Pakistan – Center for Media Engagement*. Center for Media Engagement. <https://mediaengagement.org/research/censorship-in-pakistan/>
8. Amnesty International. (2021, October 11). *Pakistan: A year of media suppression and rights abuses*. <https://www.amnesty.org/en/latest/news/2018/12/pakistan-a-year-of-media-suppression-and-rights-abuses/>
9. Robison, K. (2024, February 7). *Inside the shifting plan at Elon Musk's X to build a new team and police a platform 'so toxic it's almost unrecognizable.'* Fortune. <https://fortune.com/2024/02/06/inside-elon-musk-x-twitter-austin-content-moderation/>
10. *The Chilling: A Global Study on Online Violence against Women Journalists* | International Center for Journalists. (n.d.-c). International Center for Journalists. <https://www.icfj.org/our-work/chilling-global-study-online-violence-against-women-journalists>
11. Ibid.
12. *The Chilling: A Global Study on Online Violence against Women Journalists* | International Center for Journalists. (n.d.). International Center for Journalists. <https://www.icfj.org/our-work/chilling-global-study-online-violence-against-women-journalists>

13. Benazir Shah — Press Freedom & Advocacy - Newsroom — Coalition for Women in Journalism. (2024, June 26). Coalition for Women in Journalism. <https://www.womeninjournalism.org/threats-all/tag/Benazir+Shah#gsc.tab=0>
14. Ibid.
15. <https://digitalrightsfoundation.pk/wp-content/uploads/2017/02/Surveillance-of-Female-Journalists-in-Pakistan.pdf>
16. Irfan, A. (2023, June 12). *Censoring the Media in Pakistan*. Instick. <https://inkstickmedia.com/censoring-the-media-in-pakistan/>
17. *Gender-Based Disinformation: Advancing our understanding and response - EU DisinfoLab*. (n.d.). EU DisinfoLab. <https://www.disinfo.eu/publications/gender-based-disinformation-advancing-our-understanding-and-response/>
18. Keskin, E. (2024, February 7). *Pakistan: Surge in threats against Gharida Farooqi Deliberate attempt to silence journalist, CFWIJ and WPF Reveal* — Coalition for Women in Journalism. Coalition for Women in Journalism. <https://www.womeninjournalism.org/threats-all/pakistan-surge-in-threats-against-gharida-farooqi-deliberate-attempt-to-silence-journalist-cfwij-and-wpf-reveal>
19. Keskin, E. (2023, December 12). *Pakistan: Groundbreaking verdict for Asma Shirazi upholds media accountability in Pakistan* — Coalition for Women in Journalism. Coalition for Women in Journalism. <https://www.womeninjournalism.org/threats-all/pakistan-groundbreaking-verdict-for-asma-shirazi-upholds-media-accountability-in-pakistan>
20. *More than 60 Journalists and Media Practitioners Across Pakistan Condemn the Sexist and Derogatory Statements Made by Dr. Omer Adil Against Women in Media [Review of More than 60 Journalists and Media Practitioners Across Pakistan Condemn the Sexist and Derogatory Statements Made by Dr. Omer Adil Against Women in Media]*. Digital Rights Foundation.
21. Keskin, E. (2021b, December 24). *Pakistan: CFWIJ joins Pakistani women journalists' campaign against vicious social media attacks* — Coalition for Women in Journalism. Coalition for Women in Journalism. <https://www.womeninjournalism.org/threats-all/pakistan-cfwij-joins-pakistani-women-journalists-campaign-against-vicious-social-media-attacks>
22. https://cfj.org/wp-content/uploads/2024/09/Pakistan-Blasphemy-Report_September-2024.pdf

23. Safi, M., & Baloch, S. M. (2019, October 25). Asia Bibi arrives in Canada after leaving Pakistan. *The Guardian*. <https://www.theguardian.com/world/2019/may/08/asia-bibi-arrives-in-canada-after-leaving-pakistan>
24. Hussain, A. (2023, August 16). Mobs burn Christian churches, homes in Pakistan after blasphemy allegations. *Al Jazeera*. <https://www.aljazeera.com/news/2023/8/16/angry-mobs-burn-christian-churches-in-pakistan-after-blasphemy-allegations>
25. Amnesty International. (2024, August 20). *Pakistan: One year since Jaranwala attack, minority Christians await justice*. <https://www.amnesty.org/en/latest/news/2024/08/pakistan-one-year-since-jaranwala-attack-minority-christians-await-justice/>
26. Khuhro, Z. (2024, February 29). Pakistan: When a blasphemy accusation is evidence; the sentence often death. *Al Jazeera*. <https://www.aljazeera.com/opinions/2024/2/29/pakistan-when-a-blasphemy-accusation-is-evidence-the-sentence-often-death>
27. Ibrahim, I.. (2024, September 23). *Surge in blasphemy cases: two incidents reported in Chichawatni and Lahore*. Voicepk.net. <https://voicepk.net/2024/09/surge-in-blasphemy-cases-two-incidents-reported-in-chichawatni-and-lahore/>
28. Zakaria, R. (2017, November 24). *The borders of freedom: Blasphemy and the press in Pakistan*. CNN. <https://edition.cnn.com/2017/11/24/opinions/blasphemy-laws-pakistan-opinion-zakaria/index.html>
29. Khattab, A. (2024, September 18). CFJ report highlights devastating impact of blasphemy accusations in Pakistan. Clooney Foundation for Justice. <https://cfj.org/news/cfj-report-highlights-devastating-impact-of-blasphemy-accusations-in-pakistan/>
30. *HRFP demands justice for murder of Nazir Masih, falsely accused of blasphemy in Sargodha*. (2024, June 3). ANI News. <https://www.aninews.in/news/world/asia/hrfp-demands-justice-for-murder-of-nazir-masih-falsely-accused-of-blasphemy-in-sargodha20240603160050/#:~:text=Nazir%20Masih%20and%20his%20son,Islamic%20students%2C%20to%20attack%20them>
31. Staff Report. (2024, May 26). *44 identified as police book 500 for Sargodha mob attack on Christian man*. Pakistan Today. <https://www.pakistantoday.com.pk/2024/05/27/44-identified-as-police-book-500-for-sargodha-mob-attack-on-christian-man/>

32. Staff Report. (2024, June 5). HRCP suspects Sargodha incident was a targeted attack. DAWN.COM. <https://www.dawn.com/news/1837834>
33. *Translated and paraphrased by Digital Rights Foundation researchers.*
34. <https://digitalrightsfoundation.pk/wp-content/uploads/2017/12/UNSR-Submission-by-DRF.pdf>
35. *Addressing online violence against women and gender minorities in Pakistan.* (n.d.). World Justice Project. <https://worldjusticeproject.org/world-justice-challenge-2021/addressing-online-violence-against-women-and-gender-minorities>
36. Gossman, P. (2020, October 28). Online harassment of women in Pakistan. Human Rights Watch. <https://www.hrw.org/news/2020/10/22/online-harassment-women-pakistan>
37. *I Don't Forward Hate: an online campaign against hate speech in Pakistan.* (2019). OHCHR. <https://www.ohchr.org/en/get-involved/stories/i-dont-forward-hate-online-campaign-against-hate-speech-pakistan>
38. *beautifulsoup4.* (2024, January 17). PyPI. <https://pypi.org/project/beautifulsoup4>
39. *Pandas – Python Data Analysis Library.* (n.d.). <https://pandas.pydata.org/>
40. Crenshaw, Kimberle () "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics," University of Chicago Legal Forum: Vol. 1989: Iss. 1, Article 8.
41. IFJ survey: One in two women journalists suffer gender-based violence at work / IFJ. (2017, November 24). <https://www.ifj.org/media-centre/reports/detail/ifj-survey-one-in-two-women-journalists-suffer-gender-based-violence-at-work/category/press-releases>
42. Amnesty International. (2021a, August 13). *Pakistan: Journalists should not be harassed or intimidated through criminal justice system.* <https://www.amnesty.org/en/latest/news/2018/09/pakistan-journalists-should-not-be-harassed-or-intimidated-through-criminal-justice-system/>
43. Mulvey, Laura. (1975). Visual Pleasure and Narrative Cinema. *Screen*, 16(3), 6–18.
44. Spivak, G.C. (1988). Can the Subaltern Speak? In C. Nelson & L. Grossberg (Eds.), *Marxism and the Interpretation of Culture* (pp. 271–313). University of Illinois Press.

45. Fairclough, Norman. (1995). *Critical Discourse Analysis: The Critical Study of Language*. Longman.
46. Punjab Police District Police Officer (DPO). (2024). "TWO AHMADIYYA COMMUNITY MEMBERS SLAIN IN PHALIA" Accessed: 25 June 2024 from: <https://punjabpolice.gov.pk/node/18888>
47. Gabol. I. (2024). "Two Ahmadiyya community members slain in Phalia" Dawn.com. Available at: <https://www.dawn.com/news/1838707> (Accessed: 25 June 2024).
48. Punjab Police District Police Officer (DPO). (2024). "TWO AHMADIYYA COMMUNITY MEMBERS SLAIN IN PHALIA" Accessed: 25 June 2024 from: <https://punjabpolice.gov.pk/node/18888>
49. Hussain. A. (2024). "Local tourist killed in Pakistan's Swat over blasphemy allegations" Aljazeera.com Available at: <https://www.aljazeera.com/news/2024/6/21/local-tourist-killed-in-pakistans-swat-over-blasphemy-allegations> (Accessed: 27 June 2024).
50. A derogatory slur used in reference to members of the Ahmadiyya community. It is not only used informally, but by the government of Pakistan in official documents.
51. Digital Rights Foundation. (2023, October). *White Paper: A Southern and South-Eastern Lens on Online Harmful Content and Platform Accountability during Elections*. Digital Rights Foundation. <https://election2024.digitalrightsfoundation.pk/wp-content/uploads/2024/01/White-Paper-A-Southern-and-Southeast-Asian-lens-on-Online-Harms-Internews.pdf>
52. Council of Europe (COE). (2021). "CONTENT MODERATION: Best practices towards effective legal and procedural frameworks for self-regulatory and co-regulatory mechanisms of content moderation" Available at: <https://rm.coe.int/content-moderation-en/1680a2cc18>



DigitalRightsFoundation
"KNOW YOUR RIGHTS"



@digitalrightsfoundation



@DigitalRightsFoundation



@DigitalRightsPK



@digitalrightsfoundation



@digitalrightsfoundation



@DigitalRightsPK