Proactive Blocking through the Automated Identification of Likely Harassers

Ifat Gazia, Trevor Hubbard, Timothy Scalona, Yena Kang and Ethan Zuckerman

Abstract. Since people began interacting in computer-mediated spaces, there has been a need to block or silence abusive users. In 2014, Gamergate—a purported campaign for "ethics in game journalism," which often seemed a misogynist protest against women in computer gaming—brought the issue of online harassment to popular attention and inspired a wave of tools and techniques to mitigate online abuse. Yet, it remains a serious problem. Individuals, particularly activists and political dissidents, can face intense harassment on platforms like X (formerly Twitter), designed to silence their speech. This paper proposes a method to block likely abusers on X, using Kashmiri dissidents and Hindutva (Hindu nationalist) harassers as a case study. We first interviewed six Kashmiri dissidents who use social media for their activism to better understand their unique online experiences. Then, using a combination of text analysis and social network analysis, based on a sampling of accounts provided by the interviewees, we developed a novel filtering method. Our tests indicate that the model is 97% effective at identifying accounts that were previously blocked for harassment. This model could be useful for screening interactions on X, and preemptively filtering any it identifies as potential harassment. While it may no longer be appropriate for protecting Kashmiri users—many of whom have fled social media platforms—this model could be used in other minority communities and on other social media platforms.

1 Introduction

The global rise of social media has enabled individuals to discuss and share a variety of content, from personal stories to political issues. This networked public sphere allows voices that are often overlooked in mainstream media to mobilize, leading to influential

doi:10.54501/jots.v2i3.175

political and social movements including the Arab Spring, #MeToo, and #BlackLives-Matter. Digital platforms can amplify activists' voices in spaces that were previously inaccessible, increasing their ability to spotlight issues, organize their communities, and bring about change.

However, the affordances and infrastructure of social media can also harm vulnerable users. In 2014, campaigns like Gamergate drew widespread attention to the issues of online harassment. This online campaign targeted women in the gaming community and exposed notable women game developers and scholars to threats of doxing (making private information public), rape, and death (Valenti 2014). Although many dismissed it as a 'consumer revolt,' it is better understood as a misogynist attack on women gamers and developers designed to silence dissent (McCormick 2016).

Given the limitations of platform companies' vague content moderation strategies, this paper explores characteristics of online harassment suffered by a subset of social media users—political activists—and discusses the design of a system that could address this problem. We focus on activists from Kashmir who engage in socio-political discussion in online spaces. Social media platforms emerged as important civic spaces for marginalized Kashmiri activists around 2008 because Kashmiris enjoy few rights of assembly or freedom of expression in their homeland, which is heavily occupied by the Indian military (Human Rights Watch 2022). Online spaces became an alternative to more traditional public spheres; users engage in political discourse and vent their anger. However, well-organized Hindu nationalist trolls quickly acted to silence Kashmiri voices online. Hindutva[1] trolls operate with the tacit support of India's Bharatiya Janata Party (BJP) government and have become skilled at silencing dissent online.[2] These online attacks are accompanied by risks of physical harm, including arrest, as well as fears of being deplatformed (getting kicked off social media sites). Given the dangers of online abuse, this study explores whether it is possible to identify and block individuals who are likely to be abusive in the future.

To ground our research in the experiences and needs of the communities most affected by online abuse, we interviewed six Kashmiri activists (two based in Kashmir, four based in the United States) who engage in social and political discussion about Kashmir online, specifically on Twitter (known as X since 2023). During the online interviews in 2021, we asked the activists a series of questions about their general social media use, how they use social media to advocate for Kashmir, and the consequences of their online activism.

We experimented with creating a "harassment filter" to help limit online harassment. We first generated a dataset containing the interviewees' Twitter accounts as well as the accounts they followed and blocked (due to harassment). We used this dataset to

1. Hindutva is a recent political movement aiming to change India from a secular state to a Hindu-centric nation called the Hindu ("Defining Hindutva" 2023).
2. The popular press has widely covered the silencing of voices critical of the BJP government (see Al Jazeera 2021).

train a neural network that could accurately predict whether an account was likely to engage in harassment. This model's high performance suggests that there are underlying, detectable patterns in the behavior of accounts that make it possible to distinguish "harassers" from Kashmiri activists. The trained model could be used to preemptively filter any online interactions to identify potential harassers.

2 Literature Review

According to Lenhart, Zickuhr, and Price-Feeney (2016), 47% of US internet users have experienced online harassment or abuse. They measured direct online harassment (e.g., being called offensive names), invasion of privacy (e.g., exposure or spreading of information beyond the owner's control), denial of access (e.g., misuse of reporting tools to get a person blocked from a platform). Online harassment uses networked technologies and includes several key techniques (Marwick, Blackwell, and Lo 2016):

Doxing: revealing personal information publicly

Brigading: a group of people working together to harass an individual

Revenge porn: disseminating private photos (real or falsified) without the individual's consent

Swatting: reporting a false threat to call an emergency response team to the individual's home

Online harassment can inhibit social media participation, particularly for journalists, activists, and scholars who strive for social justice. It can therefore be understood as a civil rights and freedom of expression issue. For example, Mijatović (2016) argues that the online abuse experiences of female journalists have triggered concern and deterred free expression. Similar suppression of free speech has been examined among activists including feminist and ethnic minorities (Wright, Trott, and Jones 2020) and scholars who study sensitive topics such as racial politics and gender equality (Marwick, Blackwell, and Lo 2016).

Prior work has argued that platforms' technological affordances, governance policy, and infrastructure reinforce racist and misogynistic harassment online (Brock Jr 2020; Matamoros-Fernández 2017; Noble 2018). According to Massanari (2016), Reddit's platform design and algorithmic politics implicitly favor certain groups of users (young, white, cis-gendered heterosexual males) and marginalize others. Massanari (2016) also argues that socio-technical networks including Reddit, 4chan, Twitter, and online games have enabled toxic techno cultures to thrive. York and Zuckerman (2019) suggest that online harassment can be understood as another form of social media censorship alongside that conducted by governments and tech platforms.

To combat online harassment, major social media platforms continually adjust their definitions of harassment and police content that violates their guidelines. In 2015, Twitter expanded its conception of harassment to include prohibiting threats, targeted harassment, and the disclosure of private and confidential information (Citron 2014) to "ensure that voices are not silenced because people are afraid to speak up."[3] Many platform companies have introduced several features such as blocking content, banning, and suspending inappropriate accounts.

While platform companies seek to combat hate speech and mis/disinformation, changing policies and tools can be controversial. Platforms must strike a delicate balance between controlling toxic content and permitting unfettered speech (Goyal, Park, and Vasserman 2022). When introducing automated techniques, platforms also struggle to define what should be considered inappropriate content (Scheuerman, Branham, and Hamidi 2018). Joan Donavan (2020) points out that online platforms are generally not transparent about their decisions in moderating content. It is common to observe unjust experiences and an uproar from activists whose accounts have been taken down by platform companies for unclear reasons. For example, Carolina Are,[4] a pole dancer and activist, pointed out that Instagram's shadow ban (hiding Instagram posts from individuals who do not follow a given user) is censoring many posts for fear of being seen to promote prostitution, which threatens the survival of marginalized communities using a platform to express themselves and make a living.

2.1 Automated Content Moderation

Since human content moderation is expensive and time consuming, platforms have attempted to move to algorithmically driven semi-automated moderation systems. Since Elon Musk's acquisition of Twitter in 2022, the platform's moderation has mostly been automated. For example, Twitter developed a "qualify filter" that utilizes algorithms to detect content that average users would like to avoid seeing (Leong 2016). Facebook similarly developed word classifiers trained to detect hate speech texts and flag them for human review (Gillespie 2020; Gorwa, Binns, and Katzenbach 2020).

The heavy reliance on automation for content moderation reduces the need for manual reviews and opting for limitations on content distribution rather than outright removal of certain speech (Paul and Dang 2022), yet it has highlighted three important issues. First, automation can make it more difficult to decipher the dynamics of takedowns, raising transparency issues (Gorwa, Binns, and Katzenbach 2020; Suzor et al. 2019). Second, when moderating toxic content, content classifiers are likely to lack sufficient context and disproportionately block certain groups, thus causing representational harm to such groups (Binns et al. 2017; Gorwa, Binns, and Katzenbach 2020). Third, giving platforms the role of gatekeeping and shaping information could aggravate the asymmetrical power relationship between platforms and users (or producers), which might deter freedom of

3. https://blog.twitter.com/en_us/a/2015/policy-and-product-updates-aimed-at-combating-abuse
4. https://bloggeronpole.com/2019/07/what-instagram-pole-dance-shadowban-means-for-social-media/

expression.

In response to such problems, scholars have suggested that platforms should incorporate stakeholders (e.g., users, civil society groups) into the decision-making process (Kumar 2019; Suzor et al. 2019). For example, Gillespie (2020) suggests "designing [machine learning] tools to support the human team rather than supplant them." Prompted by these calls for collaboration between humans and automation systems, this project explores the creation of a harassment filter trained by a contextual database that reflects actual activists' experiences. We used insights from our interview process and previous designs to implement a novel approach to proactive content filtering based on machine-learning models.

We take inspiration from Tracy Chou's Block Party[5] and two other tools designed to fight online harassment. The first is Squadbox, a tool for helping recipients of email harassment coordinate a "squad" of friend moderators to shield and support them during an attack. The MIT Computer Science & Artificial Intelligence Laboratory research team designed it through a participatory design process that included "learning from harassment recipients' existing strategies to then design a tool to then design a tool to augment those strategies" (Mahar, Zhang, and Karger 2018). The second tool we consulted is Block Together—a now-defunct web application developed by volunteers to combat harassment on Twitter. Jacob Hoffman-Andrew, one of its developers, described it as a tool to defend against abuse on Twitter by blocking. It allowed Twitter users to share their list of blocked accounts that others can subscribe to (Jhaver et al. 2018). It was shut down in July 2020 because it had "gotten too big to maintain on a volunteer basis, and it's ultimately trying to fill a need that Twitter should be fulfilling on the official platform."[6]

2.2 Background: The Case of Twitter

Reporting, Muting, and Blocking. Twitter provides functions to block, mute, and report an account, which allow users to stop receiving notifications from that account (Citron 2014). Twitter defines blocking as a feature that "helps users restrict specific accounts from contacting them, seeing their Tweets, and following them."[7] Mute enables users to "remove an account's tweets from the timeline without unfollowing or blocking that account. While users will no longer receive push or SMS notifications from any muted account, muted accounts can still view the user's posts and reply to them. In this sense, muting an account is more socially delicate than blocking (Jhaver et al. 2018).

Sharing Block List (currently inactive). In 2015, Twitter introduced a feature that enabled users to share their block lists with other users, allowing users to block several accounts at once. The platform explained that it developed this tool because "some users of

5. https://www.blockpartyapp.com/about-us
6. https://blocktogether.org
7. https://help.twitter.com/en/using-x/someone-blocked-me-on-x

those high volumes of unwanted interaction on Twitter need more sophisticated tools."[8] The feature allowed users to export and share their block lists with those who are experiencing similar issues or to import another user's list and block multiple accounts at once. However, after recent updates to X, this feature is no longer available.

Third-party blocking tools (currently inactive). Activist Tracy Chou experienced extensive online abuse after organizing a study of gender ratios in the engineering teams of prominent tech companies. Seeking a more powerful alternative to shared blocklists, she created "Block Party," an "authorized agent" that could filter social media for at-risk users. For example, Block Party allowed users to preemptively block everyone who "liked" a post that wished an activist harm. Changes made to Twitter's API in summer 2023 broke Block Party and forced Chou to suspend the project.

2.3 The Kashmir Context

Since 1947, the region of Jammu and Kashmir has been under Indian control, and has experienced widespread human rights abuses and suppression (Human Rights Watch 1999). It remains one of the most heavily militarized areas in the world (Singh 2016). Individuals who dare to voice opposition to India's actions face both online and offline repression (Hassan 2020). This suppression, which has been ongoing for decades, intensified following Narendra Modi's rise to power in 2014 (Gazia 2024). In August 2019, his government revoked the region's special status, which had granted it significant autonomy and provided protections to local residents regarding residency, land ownership, and employment rights (Neuman and Frayer 2019). This move sparked concerns both within Kashmir and internationally; many feared that India intended to erode the region's unique cultural identity (Harvard Law Review 2021).

Online suppression tactics, such as mass reporting and suspension of Kashmiri Twitter accounts, have become increasingly common. Between January 2021 and December 2022, the Twitter account of the Kashmiri diaspora-led grassroots movement Stand with Kashmir (SWK) was suspended twice, along with numerous accounts belonging to Kashmiri activists and scholars (Gazia 2021). These targeted attacks on individual Twitter users mark a new phase in the digital censorship of Kashmir, which was previously characterized by frequent internet blackouts. According to the tracking website internet shutdowns.in, Kashmir has experienced over 305 shutdowns since 2012. A 4G internet ban was enforced from August 2019 to February 2021 (Jamwal 2021). Prior to 2021, Kashmiris were often unable to access the internet. Since then, targeted reporting and harassment have sought to silence Kashmiri voices less directly, but perhaps more impactfully (Gazia 2024).

8. https://help.twitter.com/en/using-x/advanced-x-block-options

3 Methods: Interviews

To understand the characteristics of online harassment experienced by social media users, particularly Kashmiri activists who engage in socio-political discussions, we interviewed six Kashmiri activists in 2021 about the challenges they experience on social media platforms, particularly Twitter, which has been very popular with Kashmiri activists. All the interviews were conducted in English language.

We developed an interview guide that sought to identify trends in social media usage, cyber-harassment on Twitter, and its consequences. After conducting six semi-structured interviews with Kashmiri activists and allies of varying degrees of visibility, we recorded and transcribed the interviews using Otter.ai and then uploaded the transcripts into Dedoose for analysis.

We used seven (often overlapping) "parent" codes to label core elements of the activity and impact of Kashmiri activism in the transcripts: Participant Background, Social Media Activity, Consequences, Harassment Description, Institutions, Kashmiri Politics, and Good Quote. "Participant background" described statements indicating the identity of the interviewee, including their work history or connection to Kashmir. "Social media activity" included descriptions of the platforms they used, the purpose for which they did so, and the amount of time spent on them. "Consequences" referred to statements describing the effect of the interviewee's social media activity and related harassment. "Harassment description" was used to code any description of cyber-harassment or physical harassment that the interviewee or their family experienced as a result of their online activism. "Institutions" was used to indicate any reference to the role of the state (whether domestic or international) or private companies in perpetuating the harassment of Kashmiri activists online. "Kashmiri Politics" was used to describe references to or descriptions of the political climate within Kashmir. Finally, "Good quote" was used to indicate a striking or powerful statement by an interviewee.

Within these parent codes, we developed multiple "child" codes to further categorize the interview data. These codes allowed us to identify the users' experience with Twitter, instances of cyber-harassment, and the individual and institutional consequences of this activity. These interviews showcased the wider harassment of Kashmiri activists online and its implications for their well-being and activism. We labeled interview participants (IP) numerically to protect their identities as IP1, IP2, IP3, IP4, IP5, and IP6.

4 Results: Interviews

The interviews highlighted clear differences between the experiences of Kashmiri activists in the diaspora and those living in Kashmir. Because the latter live in an occupied state, their Twitter posts can provoke threats or instances of physical and emotional harm.

The interviews revealed that these activists face physical intimidation, mass reporting, and doxing. The Indian government sometimes threatens the friends and families of activists who live in the diaspora to intimidate and silence them.

While Kashmiri activists living outside the region are not at direct risk of arrest, organized right-wing Hindutva trolls engage in targeted efforts to harass them on Twitter and exploit the platform's reporting function to get their accounts banned (Crawford and Gillespie 2014). These trolls send threatening direct messages and encourage their followers to falsely report the activists for alleged violations of the platform's Code of Conduct. Many interview participants described enduring mass reporting by these trolls and the resulting suspensions. Less prominent interviewees shared that while they do not experience this kind of harassment, they see it in other Twitter threads.

When asked to describe their experiences on social media and the consequences of their online activism, IP1 stated:

> "Given the current circumstances around Kashmir, the level of hate and the degree of organizing that exists in this in an Islamophobic way from a lot of Hindutva activists, I think that the reason I keep getting flagged is in a targeted way by either actual people that go out there and troll Twitter accounts or by bots that may have been created that look for my accounts like mine, to flag them."

IP2, responding to the same question, said:

> "The harassing they did wasn't enough, so they decided to dox me. There was this guy on Twitter with almost 100,000 followers who basically researched me and put all my personal information on Twitter. And then I received so many misogynist and very bad comments, basically, which were not politically harmful to me. But personally, they affected me so much. The guy had made sure to block me before he doxed me. But of course, like he had 100,000 followers, people saw it. And my friends tried not to send me the screenshots of it, so I had to log into a different account to see what was really happening."

Other interviewees reported that the Indian government directly targets the friends and family of activists living in the diaspora as a form of intimidation. These activists face a mixture of targeted harassment, doxing, and threats while their families are at risk of arrest and physical harm. IP1 noted that they experience significantly more harassment for their Kashmiri activism than for their advocacy on other sensitive topics including Palestinian liberation and gun violence in the United States. However, IP2 and IP3 reported that some Indian allies support and share their Kashmiri activism on Twitter to break the cycle of propaganda and educate their peers.

4.1 The Solution

As X continues to silence dissident voices and succumb to government demands (Mehrotra and Menn 2023), IP2 and IP5 proposed that it should create transparent accountability and suspension mechanisms, improve communication with affected groups, and prevent the automated banning of mass-reported accounts. IP2 explained that human moderators need to be versed in and sensitive to social justice issues, including the circumstances regarding the tensions between Hindu nationalists and Kashmiris. Every interview participant emphasized the need for the company to promote freedom of speech and provide a safe platform for marginalized voices to share their experiences. However, it seems unlikely since 80% of the engineers who work on trust and safety have been terminated since Elon Musk's acquisition of the platform and the outsourced content moderation teams have been downsized (Robison 2024).

4.2 Consequences

The organized and targeted harassment described by the interviewees silences and suppresses Kashmiri voices. The participants reported four main consequences of this cyber-harassment: 1) the threat of deplatforming, 2) the fear of physical or emotional harm, leading to self-censorship, 3) the loss of social capital, and 4) distrust of X. Many activists have been falsely banned multiple times as a result of mass reporting by Hindu nationalists on Twitter; IP1, IP2, and IP3 described how this happened to them and jeopardized their social connections and activism. They also described the deep emotional consequences of contending with Twitter's automated banning system and their fear of losing contact with their online community. Media activism has become a high-risk activity: with every tweet, they risk deplatforming, arrest, and physical harm to themselves or their family.

All six interviewees noted that they use pseudonyms and censor their content to avoid harassment from Hindu nationalist trolls and harm from the Indian government. Other activists, like IP3, have silenced their activity to protect loved ones and limit the harassment, which erodes online discourse about Kashmiri issues. IP1 and IP2 reported that they had spent years building their social networks for their activism and that the loss of these networks would hinder their work. Finally, IP1, IP3, and IP4 described losing trust in X for what they see as the platform's complicity in, and perpetuation of, the cyber-harassment of Kashmiri activists; they have considered deleting their accounts. Only one of our interviewees, IP2, still has an active account on X. They have not posted about Kashmir for 3 years, resulting in a lack of recent tweets on the subject.

4.3 Institutional Complicity

Interviewees blamed both Twitter and the Indian government for silencing Kashmiri activists, noting an intersection between their actions and inactions. IP1, IP2, and IP3 recounted that when activists were banned due to mass reporting from Hindu nationalist

trolls, Twitter often failed to respond to their attempts to reinstate their accounts and gave irrelevant reasons for their suspension from the platform. The platform required all three activists to follow up about their suspensions, such as providing a driver's license to verify their identity. This made all three interviewees feel unsafe and they feared retaliation from the Indian government if they were to do so. IP2 and IP3 suggested that Twitter's financial investments in India might be influencing their explicit suppression of Kashmiri activists' accounts. They both reported feeling as if Twitter was complicit in this activity because of the disproportionate blocking of Kashmiri activists on the platform and the failure to suspend right-wing Hindutva trolls, including accounts that "openly congratulate" their followers for mass reporting an account and getting it suspended.

IP1, IP3, and IP5 cited American values such as freedom of speech to explain their disappointment with Twitter's suppression of Kashmiri voices. Some claimed they would somewhat understand the company's censorship if Twitter operated in a country with policies that promoted such activities. However, Twitter's behavior contradicts American principles.

5 Methods: Codesign, Informed by Interviews

Informed by the interviews with activists and inspired by the "shared block lists" tactics used by feminist activists like Chou, we built a machine-learning-based "harassment filter." Our filter learns from known harassers to detect accounts that have been identified as likely harassers, which the user can then choose to block. Each activist could "teach" the model the behaviors of their harassers by providing a list of accounts she had been forced to block online. The model analyzes linguistic and network features of the harassers' accounts (i.e., the account's tweets/likes/retweets, as well as its follower/following networks) to identify "patterns of harassment." These patterns can then be used to identify new harassers.

We framed our problem as a binary classification problem: when provided with information from a particular Twitter account in the training set, could the model identify whether that account had taken part in any harassment of Kashmiri activists? If a model performed accurately on this task, then it could be used to determine whether accounts outside of the training set were harassers or not."

Most previous attempts at algorithmic content moderation have been focused on textual content (Fortuna and Nunes 2018). A variety of Neuro Linguistic Programming (NLP) techniques have been used to classify whether particular excerpts of text could be considered abusive or hateful. More recent research has shown that incorporating network-based features can improve the accuracy of these abuse detection models (Chopra et al. 2020; Cécillon et al. 2019, 2021). Graph-embedding techniques have made it easier to integrate these graph-based features, which aim to create representations of

graph structures in a vector space (Goyal and Ferrara 2018).

When building our "harassment filter," we took inspiration from both the content- and graph-based content moderation approaches. We trained a multi-layer perceptron on both textual data (word2vec embeddings of the tweets and retweets of various accounts in the English language) and network data (node2vec embeddings of the follower/following graphs of accounts), aiming to classify accounts as either "blocked" or "not blocked." Our best-performing model achieved an F1 score of 96.8%.

5.1 Data Collection

"Harassment" is generally difficult to define, and can take on a variety of forms depending on the context. Instead of rigidly defining the concept, we relied on the lived experience of Kashmiri activists who had experienced online harassment. We identified 62 accounts that SWK's Twitter account had blocked, managed by one of the interview participants, due to their harassment.

To complement this group, we sought to identify Twitter accounts that did not exhibit the behaviors of the blocked accounts to teach the neural network what the behaviors of a "not-blocked" account looked like. An interviewee developed a list of 62 not-blocked accounts linked to Kashmiri activists/sympathetic to the cause for this purpose.

To bolster our dataset, we identified an additional 67 accounts that had been closely connected to these blocked and not-blocked accounts (through follower/following relationships). Our lead author, who has been a direct victim of Hindutva harassment, read through a sample of each account's tweets within the previous month to assess whether a Kashmiri activist would be likely to block them. She characterized the tone of the language used in the tweets and their content to decide whether their behaviors could be considered "harassment" and thus determine how the harassment filter ought to label the account. For example, an account that referred to Muslim Kashmiri dissidents as 'Jihadis' or 'terrorists' would be categorized as "blocked."

This process generated a total of 191 accounts (86 blocked, 105 not blocked). While this appears to be a small data set, it represents a large number of active Kashmiri activists and known trolls at the time it was generated. We used Twitter API to scrape the following content to analyze the behaviors of each account:

- A list of the users the account follows
- A list of the users that were following the account
- The account's most recent 450 tweets/retweets
- The account's most recent 600 liked tweets

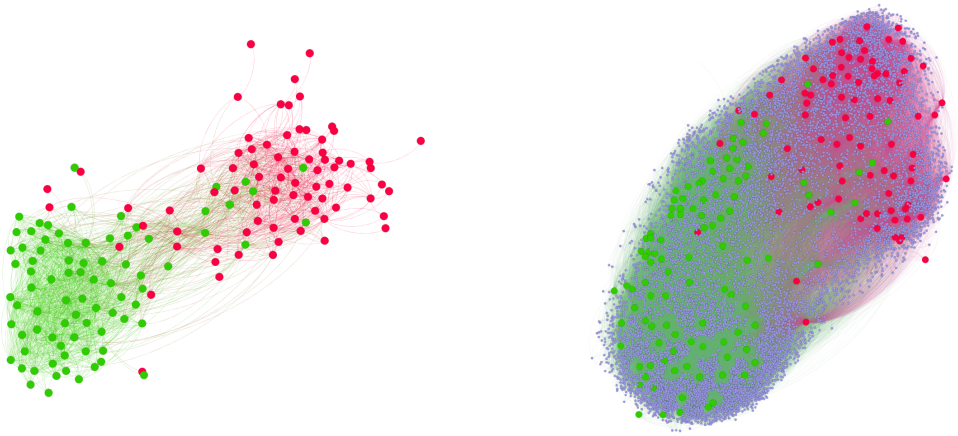
Since tweets/retweets are returned from the same API endpoint, different accounts

had varying numbers of tweets vs. retweets. The data that was collected from each account generally fit into one of two categories: social network data or textual data. Analyzing image/video content shared by any of these accounts was beyond the scope of the project.

5.2 Social Network Data Pre-Processing

Using the follower/following lists we scraped from the accounts of interest, we created a large directed network. For instance, if Account A was following Account B, we draw an edge from Account A's node to Account B's node. This network included 477,631 nodes and 915,873 edges. Most of these nodes represented "adjacent" accounts, which were not initially identified as blocked/not blocked but were connected to those accounts through follower/following relationships.

For ease of analysis, Figure 1b only includes adjacent nodes that were connected to 10 or more blocked or not-blocked accounts, which reduces the sample to 8,839 nodes, connected by 161,341 edges. Figure 1b depicts a clear separation between the blocked and not-blocked subgraphs.



(a) (b)

Figure 1: Visualizations of the graphs that were generated. (a) contains only the blocked (red) and not-blocked accounts (green), while (b) also contains the adjacent accounts (purple).

5.3 Textual Data Pre-processing

The accounts contained a total of 167,038 unique tweets broken down into the following types:

- 43,017 tweets directly from accounts
- 29,027 retweets
- 94,924 liked tweets

These tweets represented a mixture of English, Hindi, and Kashmiri words written in English text. While typical NLP analyses usually involve lemmatization of text as an important pre-processing step, we omitted this step due to the difficulties of doing so using a multilingual corpus.

Other pre-processing steps included:

- Sentence tokenization
- Removing punctuation
- Removing English stopwords
- Removing hyperlinks
- Removing mentions of other users (i.e., "@username")

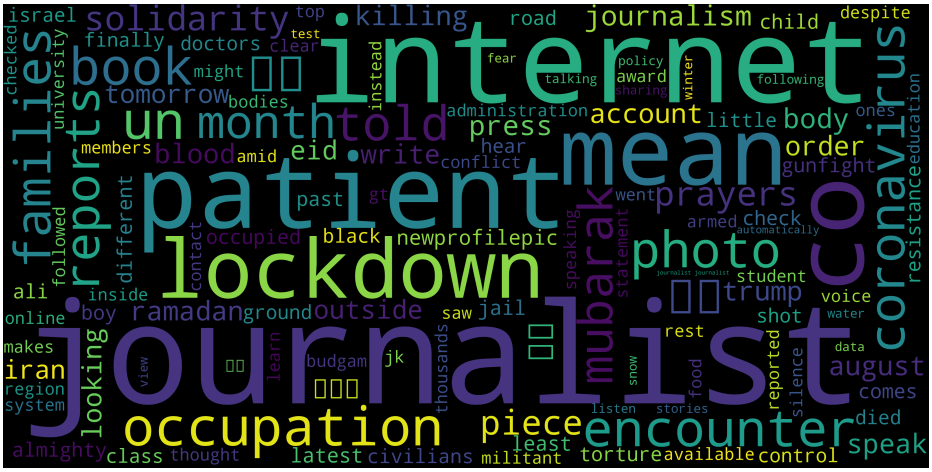
We then listed the most frequently occurring words from the English language tweets in blocked and not-blocked accounts and created word clouds of each group's most frequently occurring words that did not appear in the other group's list (see Figure 2 on the following page). These serve as a straightforward sanity check: unblocked accounts use words that talk about the "occupation" of Kashmir and police "lockdowns," while blocked accounts talk about "terrorists" and frequently use the Hindu honorific "-ji" in referring to individuals.

5.4 Experiments

In order to train a model using this data, we needed to transform it into a representation that was easier to work with. We used a word2vec model to learn a 128-dimensional vector representation of each token appearing in the corpus (Mikolov et al. 2013). word2vec is one of many algorithms designed to produce "word embeddings" from corpuses of text; previous attempts at classifying abusive language using this method have been successful (Fortuna and Nunes 2018). For the social network data, we used a node2vec model to learn 128-dimensional vector representations for each account node (Grover and Leskovec 2016). While more sophisticated models are available for both word embedding and node embedding, we chose well-understood methods because we are evaluating a linguistically complex data set.

We used these vector representations to create four classifiers in an attempt to find the best model configuration:

Text only: a multilayer perceptron trained only on the word2vec embeddings. For each blocked/not-blocked account, we created an "account-level" word embedding, which was a weighted average of all of the word embeddings that correspond to that account's textual data



(a)

(b)

Figure 2: Word clouds showing the words most frequently used by the (a) not-blocked accounts that were not frequently used by the blocked accounts and (b) blocked accounts that were not frequently used by the not-blocked accounts.

Network only: a multilayer perceptron trained only on the node2vec embeddings of each blocked/not-blocked account

Bagged: an ensemble model that averaged the scores of the previous two models to produce a new score

Split: the final output layers of the text-only and network-only models were removed, and the outputs of each model were concatenated together and fed into another multilayer perceptron

Due to the imbalance of blocked and not-blocked accounts, we randomly removed the latter from the train/test pool until we had an equal number in each group. We then trained each model on 70% of the available data. The code underlying this training process is available in our team's GitHub repository.[9]

6 Analysis: Blocking Tool

We performed 4-fold cross-validation to test the robustness of our models. Average F1 scores (a weighted average of precision and recall) for each model across all four cross-validation folds are presented below.

Table 1: F1 scores per model type.

Text Only	Network Only	Bagged	Split
0.887	0.951	0.955	0.968

These results suggest that there are underlying patterns across multiple modalities (both text content and network dynamics) that distinguish the X accounts of Kashmiri activists from those they have blocked. Network data seemed to provide a stronger signal than the textual data, but combining the two modalities produced the most accurate models. Using this model, one could conceivably screen interactions from new, previously unseen accounts. Since the model is trained to distinguish harassers from non-harassers, it could intervene (via blocking the user or programmatically hiding their replies) if it detected account behavior that was aligned with the harassers' behaviors.

7 Limitations and Future Work

Despite their promise, the high accuracy of our models is not entirely surprising considering the smaller size of our dataset. To better understand the models' performance, we would need to increase the size of our dataset (either through further manual labeling of the adjacent accounts or further crowdsourced consolidation of other activists' block

9. https://github.com/trevbook/Harassment-Filter

lists). Further experiments on the models' architecture could also be helpful, as it may be that either network- or text-based models are less effective with larger data sets and require different methods of combination.

When we initially conceived of the idea of a "harassment filter," we envisioned using it in concert with the Twitter API to automatically filter replies to tweets. Twitter's API allows users to programmatically hide other users' replies;[10] one could imagine using this model to automatically hide any replies from accounts classified as potential harassers. Much more development would be necessary to enable the models' use online, as our current conception uses a fixed data set.

Unfortunately, changes made to the Twitter API's pricing structure since Musk's takeover have made the operation of such a model unfeasibly expensive. Once-free API features are now prohibitively expensive. Other anti-harassment projects, like Block Party, have been placed on indefinite hiatus because of these changes. One can only hope that the pricing structure becomes more accessible in the future to permit further grassroots experimentation on anti-harassment strategies.

The abandonment of Twitter by many Kashmiri activists is perhaps more striking. Of the six activists we interviewed, only one has remained on the platform. More broadly, there is significantly less conversation on Twitter from known activists, perhaps as a result of the sustained harassment campaigns. We hope our proposed proactive blocking method will be useful for communities experiencing sustained harassment.

10. https://developer.twitter.com/en/docs/twitter-api/tweets/hide-replies/introduction

References

Al Jazeera. 2021. "India's Cyber Volunteers: Surveillance, Censorship, and the Internet," November. https://www.aljazeera.com/news/2021/11/29/india-cyber-volunteers-surveillance-censorship-internet-social-media.

Binns, Reuben, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. "Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation." In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II,* 405–15. Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-319-67256-4_32.

Brock Jr, André. 2020. *Distributed Blackness: African American Cybercultures.* NYU Press.

Cécillon, N., V. Labatut, R. Dufour, and G. Linarès. 2019. "Abusive Language Detection in Online Conversations by Combining Content-and Graph-Based Features." *Frontiers in Big Data* 2. https://doi.org/10.3389/fdata.2019.00008.

———. 2021. "Graph Embeddings for Abusive Language Detection." *SN Computer Science* 2 (1): 37. https://doi.org/https://doi.org/10.1007/s42979-020-00413-7.

Chopra, S., R. S. Sawhney, P. Mathur, and R. R. Shah. 2020. "Hindi-English Hate Speech Detection: Author Profiling, Debiasing, and Practical Perspectives." In *Proceedings of the AAAI Conference on Artificial Intelligence,* 34:386–93. 01. https://doi.org/https://doi.org/10.1609/aaai.v34i01.5374.

Citron, D. K. 2014. "Addressing Cyber Harassment: An Overview of Hate Crimes in Cyberspace." *Case W. Res. JL Tech. & Internet* 6 (1). https://heinonline.org/HOL/LandingPage?handle=hein.journals/caswestres6&div=4&id=&page=.

Crawford, K., and T. Gillespie. 2014. "What is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint." *New Media & Society,* https://doi.org/https://doi.org/10.1177/14614448145431.

"Defining Hindutva." 2023. https://www.hindutvaharassmentfieldmanual.org/defininghindutva.

Donavan, J. 2020. "Why Social Media Can't Keep Moderating Content in the Shadows." *MIT Technology Review* (November 6, 2020). https://www.technologyreview.com/2020/11/06/1011769/social-media-moderation-transparency-censorship/.

Fortuna, P., and S. Nunes. 2018. "A Survey on Automatic Detection of Hate Speech in Text." *ACM Computing Surveys* 51, no. 4 (July 31, 2018): 1–30. https://doi.org/10.1145/3232676.

Gazia, Ifat. 2021. "Silicon Valley Must not Silence Kashmir." *Tech Policy Press* (May 24, 2021). https://www.techpolicy.press/silicon-valley-must-not-silence-kashmir/.

Gazia, Ifat. 2024. “Silencing Kashmir: The Struggle for Voice in the Face of Social Media Censorship.” Tech Policy Press, March 7, 2024. https://www.techpolicy.press/silencing-kashmir-the-struggle-for-voice-in-the-face-of-social-media-censorship/.

Gillespie, T. 2020. “Content Moderation, AI, and the Question of Scale.” *Big Data & Society* 7 (2): 205395172094323. https://doi.org/10.1177/205395172094323.

Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. “Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance.” *Big Data & Society* 7 (1): 2053951719897945. https://doi.org/https://doi.org/10.1177/2053951719897945.

Goyal, Nitesh, Leslie Park, and Lucy Vasserman. 2022. ““You Have to Prove the Threat is Real”’: Understanding the Needs of Female Journalists and Activists to Document and Report Online Harassment.” In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems,* 1–17. https://doi.org/https://doi.org/10.1145/3491102.3517517.

Goyal, Palash, and Emilio Ferrara. 2018. “Graph Embedding Techniques, Applications, and Performance: A Survey.” *Knowledge-Based Systems* 151:78–94. https://doi.org/https://doi.org/10.1016/j.knosys.2018.03.022.

Grover, Aditya, and Jure Leskovec. 2016. “node2vec: Scalable Feature Learning for Networks.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 855–64. https://doi.org/https://doi.org/10.1145/2939672.2939754.

Harvard Law Review. 2021. “From Domicile to Dominion: India’s Settler Colonial Agenda in Kashmir.” *Harvard Law Review* (May 10, 2021). https://harvardlawreview.org/print/vol-134/from-domicile-to-dominion-indias-settler-colonial-agenda-in-kashmir/.

Hassan, Ali. 2020. ““My Phone Haunts Me”: Kashmiris Interrogated and Tortured by Cyber Police for Tweeting.” The Intercept, December 6, 2020. https://theintercept.com/2020/12/06/kashmir-social-media-police/.

Human Rights Watch. 1999. *Behind the Kashmir Conflict - Background.* Human Rights Watch Report. https://www.hrw.org/reports/1999/kashmir.

———. 2022. “India: Repression Persists in Jammu and Kashmir.” Human Rights Watch, August 2, 2022. https://www.hrw.org/news/2022/08/02/india-repression-persists-jammu-and-kashmir.

Jamwal, A. 2021. “4G Is Back in J&K After 18 Months, But it Can’t Compensate for What We Lost.” *The Wire,* https://thewire.in/rights/jammu-and-kashmir-4g-internet-costs.

Jhaver, Shagun, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. "Online Harassment and Content Moderation: The Case of Blocklists." *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, no. 2 (March 21, 2018): 1–33. https://doi.org/https://doi.org/10.1145/3185593.

Kumar, Sangeet. 2019. "The Algorithmic Dance: YouTube's Adpocalypse and the Gatekeeping of Cultural Content on Digital Platforms." *Internet Policy Review* 8 (2): 1–21. https://doi.org/doi:10.14763/2019.2.1417.

Lenhart, Amanda, Kathryn Zickuhr, and Mary Price-Feeney. 2016. *Online Harassment, Digital Abuse, and Cyberstalking in America.* Report. Data & Society Research Institute, November 21, 2016. https://www.datasociety.net/pubs/oh/Online_Harassment_2016.pdf.

Leong, Emil. 2016. "New Ways to Control your Experience on Twitter." *Twitter Product Blog,* https://blog.x.com/en_us/a/2016/new-ways-to-control-your-experience-on-twitter.

Mahar, Kaitlin, Amy X Zhang, and David Karger. 2018. "Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems,* 1–13. https://doi.org/https://doi.org/10.1145/3173574.3174160.

Marwick, Alice E., Lindsay Blackwell, and Karina Lo. 2016. *Best Practices for Conducting Risky Research and Protecting Yourself from Online Harassment.* Data and Society Institute. https://datasociety.net/pubs/res/Best_Practices_for_Conducting_Risky_Research-Oct-2016.pdf.

Massanari, Adrienne. 2016. "#Gamergate and The Fappening: How Reddit's Algorithm, Governance, and Culture Support Toxic Technocultures." *New Media & Society* 19 (3): 329–46. https://doi.org/10.1177/1461444815608807.

Matamoros-Fernández, Ariadna. 2017. "Platformed Racism: The Mediation and Circulation of an Australian Race-based Controversy on Twitter, Facebook and YouTube." *Information, Communication & Society* 20 (6): 930–46. https://doi.org/https://doi.org/10.1080/1369118X.2017.1293130.

McCormick, Rich. 2016. "Gamergate: A Misogynist Harassment Campaign Disguised as Consumer Revolt." *The Verge* (March 12, 2016). https://www.theverge.com/2014/11/4/7153549/gamergate-campaign-video-game-ethics-feminism-harassment.

Mehrotra, K., and J. Menn. 2023. "How India Tamed Twitter and Set a Global Standard for Online Censorship." *Washington Post* (November 8, 2023). https://www.washingtonpost.com/world/2023/11/08/india-twitter-online-censorship/.

Mijatović, Dunja. 2016. "New Challenges to Freedom of Expression: Countering Online Abuse of Female Journalists." *Organization for Security and Co-operation in Europe* (February 4, 2016). https://www.osce.org/fom/220411.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *arXiv [cs.CL],* https://doi.org/https://doi.org/10.48550/arXiv.1301.3781.

Neuman, Scott, and Lauren Frayer. 2019. "In Unprecedented Move, India Revokes Kashmir's Special Status, Sparks Fears Of Unrest." *NPR* (August 5, 2019). https://www.npr.org/2019/08/05/748170695/in-unprecedented-move-india-revokes-kashmirs-special-status-sparks-fears-of-unre.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press. ISBN: 9781479837243.

Paul, Ken, and Sheila Dang. 2022. *Exclusive: Twitter Leans on Automation to Moderate Content as Harmful Speech Surges,* December 3, 2022. https://www.reuters.com/technology/twitter-exec-says-moving-fast-moderation-harmful-content-surges-2022-12-03/.

Robison, K. 2024. "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* (February 6, 2024). https://fortune.com/2024/02/06/inside-elon-musk-x-twitter-austin-content-moderation/.

Scheuerman, Monica K., Stacy M. Branham, and Farzaneh Hamidi. 2018. "Safe Spaces and Safe Places: Unpacking Technology-Mediated Experiences of Safety and Harm with Transgender People." *Proceedings of the ACM on Human-computer Interaction* 2 (CSCW): 1–27. https://doi.org/https://doi.org/10.1145/3274424.

Singh, Ranjit. 2016. "Kashmir: The World's Most Militarized Zone, Violence After Years of Comparative Calm." *Forbes* (July 12, 2016). https://www.forbes.com/sites/ranisingh/2016/07/12/kashmir-in-the-worlds-most-militarized-zone-violence-after-years-of-comparative-calm/?sh=268755223124.

Suzor, Nicolas P., Shannon M. West, Ashley Quodling, and Jillian York. 2019. "What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation." *International Journal of Communication* 13:18. https://ijoc.org/index.php/ijoc/article/view/9736.

Valenti, Jessica. 2014. "Gamergate is Loud, Dangerous and a Last Grasp at Cultural Dominance by Angry White Men." *The Guardian* (October 21, 2014). https://www.theguardian.com/commentisfree/2014/oct/21/gamergate-angry-men-harassing-women.

Wright, Scott, Victoria Trott, and Christopher Jones. 2020. "'The Pussy Ain't Worth it, Bro': Assessing the Discourse and Structure of MGTOW." *Information, Communication & Society* 23 (6): 908–25. https://doi.org/10.1080/1369118X.2020.1751867.

York, Jillian, and Ethan Zuckerman. 2019. "Moderating the Public Sphere." *Human rights in the Age of Platforms* 137:143. https://library.oapen.org/bitstream/handle/20.500.12657/24492/1/1005622.pdf#page=184.

Authors

Ifat Gazia (igazia@umass.edu) is a Ph.D. Candidate at the Department of Communication at University of Massachusetts Amherst.

Trevor Hubbard (trevormhubbard@gmail.com) is an independent researcher.

Timothy Scalona (Tscalona12@gmail.com) is a J.D. Candidate at Suffolk University Law School and Master's in Public Policy at University of Massachusetts Amherst.

Yena Kang (ykang@umass.edu) is a Ph.D. Candidate at the Department of Communication at University of Massachusetts Amherst.

Ethan Zuckerman (ethanz@umass.edu) is a professor at the Department of Computer Science, Public Policy and Communication at University of Massachusetts Amherst.

Acknowledgements

We express our gratitude to Professor Ethan Zuckerman for his invaluable encouragement and support in bringing this paper to fruition. Professor Zuckerman recognized the potential of this work from its inception as a class project for his Fixing Social Media course in 2021. Additionally, we extend our appreciation to the activists who spoke with us and those who shared their block lists with us and made this project possible. Furthermore, we would like to acknowledge this journal for giving us the opportunity to publish this research.

Data availability statement

The interview appendix has been submitted as a separate work file. All the pictures used in the paper as also attached as separate files. The code is available on GitHub: https://github.com/trevbook/Harassment-Filter. The .bix text file for references as well as our response to journal comments are attached as separate files.

Funding statement

We received no financial compensation from any institution or agency for undertaking this work.

Ethical standards

Not Applicable

Keywords

Hindutva; online harassment; content moderation; Kashmir; activism; dissidents; blocking