Online Harassment and Blocklists

Shagun Jhaver, Sucheta Ghoshal, Eric Gilbert, Amy Bruckman

Georgia Institute of Technology

Online Harassment

- A growing concern for many social media platforms.
- It involves undesirable behaviors such as
  - revealing sensitive information about someone online
  - posting threats of violence, and
  - committing technological attacks.
- Victims of online harassment suffer from anxiety and distress.
- Coordinated harassment campaigns
  - Such campaigns are organized by hate groups
  - They overwhelm targets by synchronously flooding their social media feeds
  - One group that has recently gained much attention in the popular media for such campaigns is **GamerGate**.

Online Moderation Tools and their Limitations

Current social media platforms use a variety of technical and design approaches to police their content and prevent abuse:

- ***Distributed Social Moderation***: On sites like Facebook, Reddit and Yik Yak, the content is moderated through a voting mechanism where registered users *upvote* or *downvote* each each submission or comment.
- ***Centralized moderation***: On sites like Facebook groups and Reddit forums (subreddits), a small number of users called moderators manually remove abusive posts.
- ***Flagging and reporting systems***: On sites like Twitter, systems for flagging and reporting abusive content are offered. Flagging systems often rely on *terms of use*

Although the above mentioned approaches are widely used, they suffer from a number of shortcomings:

- In the *distributed social moderation model*, the abuse victims have to look at the gruesome images, death threats, slur laden posts, etc. in order to report the relevant posts, and this re-victimizes them.
- In *the centralized approach*, a few moderators have to spend countless hours in order to maintain the community.
- The *flagging mechanisms for reportin*g offensive content have limited vocabulary and leave little room for articulation of disagreements about what is offensive or acceptable

Twitter Blocklists

- Blocklists are third party Twitter applications that extend the basic functionality of individual blocking by letting users quickly block all accounts on a community-curated or algorithmically generated list of block-worthy accounts.
- Everyone has slightly different boundaries, and the use of blocklists can provide users an experience which is more customized to their needs.
- Blocklists are used to:
  - Address online abuse
  - Filter spam
  - Help ISIS radicalization

Methods

- Participant Sampling: We employed two separate groups as our study sites: the first group is composed of people who were added to GGBL, a GamerGate specific blocklist, and the second group contains people who subscribed to GGBL
- Interviews: We invited the sampled users to participate in semi-structured interviews with us by contacting them on Twitter. We conducted *28 interviews*.

Findings: Online Harassment

- ***What defines online harassment?*** Our participants had different perceptions of online harassment: they characterized as online harassment acts ranging from someone posting a spoiler about the new Star Wars movie to someone sending them death threats.
- ***Who are the harassers?*** Participant accounts suggest that prominent perpetrators of harassment include groups ranging from GamerGate supporters and GamerGate opponents to trolls, Bernie Sanders fans and radical feminist groups.
- ***Tactics used by harassers:*** Subtle threats, dogpiling, identity deception, brigading hashtags, sealioning, doxing
- ***Who is vulnerable to harassment***: Transgender community, Muslim users, women of color, feminists

Findings: Moderation on Social Media

"Why isn't there just a filter for eliminating some random person I've never interacted with yelling racist slurs at me?" (Interview subject BS-02)

- Moderation is politically driven.
- Platforms lack interest in developing anti-abuse tools.
- Ineffective enforcement of anti-harassment policies.
- Platforms design affects moderation.
- Users leave SNS because of ineffective moderation.
- Strategies used to prevent online abuse:
  - Using multiple accounts
  - Hiding posts from greater visibility
  - Deciding when to engage
  - Getting help from law enforcement
  - Avoiding spaces where abuse is likely

Findings: Blocklists

Why do users subscribe to blocklists?
- They suffer harassment
- As a preemptive measure
- Disillusion with a group

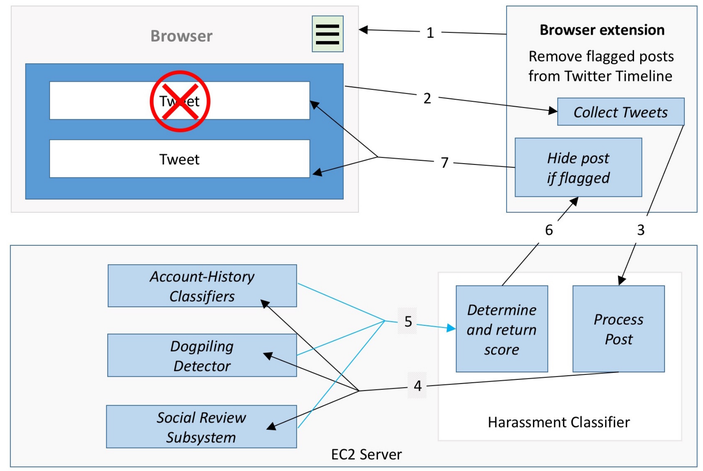
How did user experience change after using blocklists?
- Noticed false positives
- Some harassment still occurred

Algorithmically v/s Socially curated blocklists:
- Complexity of decisions
- Objectivity
- False positives

Blocklists block too much/block unfairly
- Some users suffered professionally
- Over blocking can encourage more aggressive behavior
- Lack of transparency

Appeals Procedure
- Too much effort required to appeal
- Inefficient
- Doxing

Future Work



DI1: Design to use a hate speech classifier.
DI2: Design to improve precision in moderation.
DI3: Allow user control over degree of caution in moderation (sliders).
DI4: Design to control dogpiling.
DI5: Design to make appeals accessible and efficient.
DI6: Leverage account metadata in filtering.
DI7: Design to combine algorithmic and social curation.