

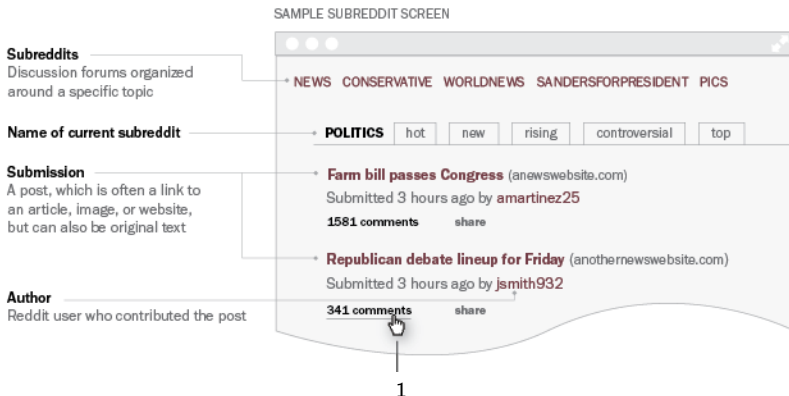
# Moderation Suite for Reddit

Himanshu Singh

May 24, 2023

# Problem Domain

## Parts of a Reddit screen



<sup>1</sup><https://www.pewresearch.org/journalism/2016/02/25/seven-in-ten-reddit-users-get-news-on-the-site/>


# Problem Description

Posted by u/IProposeThis 1 year ago 🗨️ 👍 6 🗨️ 3 🗨️ 5 🗨️ 4 🗨️ 2 🗨️ 2


4.6k

Found screenshots of the Israeli propaganda app ordering recruits to brigade and vote manipulate Reddit posts. They are fucking up Reddit!


3/6




RETWEET [this tweet](#) by Honest Reporting to get the Guardian to correct their article.




LIKE and COMMENT on [Angelina's tweet](#) to set the record straight.




UPVOTE [this post](#) on the reddit page r/worldnews to keep people around the world informed.



LIKE [Amir's comment](#) to criticize the false headline about Palestinian woman and baby who were allegedly killed by the IDF.



SHARE [this post](#) by HonestReporting to get outlets such as VOX to change their headlines.



Reuters refuses to report about all the violence directed at Israel. LIKE [Netta's comment](#) to highlight the unfair and one-sided characterization of Israel.

2

<sup>2</sup>[https://www.reddit.com/r/israelexposed/comments/n4etue/found\\_screenshots\\_of\\_the\\_israeli\\_propaganda\\_app/](https://www.reddit.com/r/israelexposed/comments/n4etue/found_screenshots_of_the_israeli_propaganda_app/)

## Problem Description



Ethics\_Woodchuck · 5 yr. ago

Here are the Chat Logs from the "burgersandfries" channel that led to Gamergate.

<https://puu.sh/boAEC/f072f259b6.txt>

A coupe of examples from the logs.

Aug 21 17.49.48 <rd0951> ./v should be in charge of the gaming journalism aspect of it.  
/pol should be in charge of the feminism aspect, and /b should be in charge of  
harassing her into killing herself

Aug 27 10.12.46 <Jiakki> so what are your guys' thoughts on feminism? Aug 27 10.12.57  
<Drinky\_Kraw> poisonous marxist scum, kill it

You didn't need a brigade, these kind of people literally created the movement. Breitbart didn't plot to take over anything, they just saw a receptive audience already sharing a similar ideology. It doesn't take a conspiracy to predict that a movement started on 4chan with heavy /pol involvement might end up pushing alt-right propaganda.



64



Give Award

Share

Report

Save

<sup>3</sup><https://www.reddit.com/r/SubredditDrama/comments/5z4o39/comment/devwkh2/>

## Problem Description



[deleted] · 1 yr. ago

lmao

"We banned them for brigading 80 times in a month. 79 times is ok, but 80 is just taking it too far!"



727



Share

Report

Save

4



NomaiTraveler · 1 yr. ago

being furry is zoophile blackface

I think brigading is most identifiable when people from a specific subreddit are flooding into a different subreddit across a variety of posts (not just a single crosspost)



100



Give Award

Share

Report

Save

5

---

<sup>4</sup><https://www.reddit.com/r/SubredditDrama/comments/pfz0d2/comment/hb7qlhp/>

<sup>5</sup><https://www.reddit.com/r/SubredditDrama/comments/pfz0d2/comment/hb7xd82/>

# In-built Measures

- Spam Filters

**spam** - This will remove the item and mark it as spam. The spam filter, in turn, will take notice of this and slowly try to learn from spammed items.

*Content removed by the spam filter used to be listed in the moderation queue for review, but now it goes straight into the spam queue.*

6

- Crowd Control

## CROWD CONTROL

Automatically collapse or filter content from people who aren't trusted users within your community yet.

### Comments

**Off:** Use Crowd Control to automatically collapse comments from users you're not sure about.



### Hold Crowd Controlled comments for review NEW

Instead of collapsing crowd controlled comments will be held for review in mod queue.



### Posts NEW

**Off:** Use Crowd Control to automatically filter posts from users you're not sure about.



7

<sup>6</sup><https://mods.reddithelp.com/hc/en-us/articles/360010090132>

<sup>7</sup><https://www.reddit.com/r/{subreddit}/about/edit?page=safety>

# Bots

- AutoModerator

Things it can't do:

- Re-check content except when something has been reported or when something is edited
- Detect reposts
- Make decisions based on vote score
- Make decisions based on any piece of content other than the submission or comment that is currently being examined (The only exception is that the properties of the parent submission may be considered when evaluating comments.)

8

- SaferBot

[Send a modmail to the /r/Saferbot team](#) with a request to bring your community under Saferbot's aegis.

9

- Flair\_Helper

Flair\_Helper has run up against a recent limitation placed by the reddit admins upon mods, namely that you cannot view your aggregate modlog if you mod more than 500 subreddits. For this reason Flair\_Helper has been down all of today.

10

---

<sup>8</sup>[https://www.reddit.com/r/AutoModerator/wiki/no\\_can\\_do/](https://www.reddit.com/r/AutoModerator/wiki/no_can_do/)

<sup>9</sup><https://www.reddit.com/r/Saferbot/wiki/introduction/>

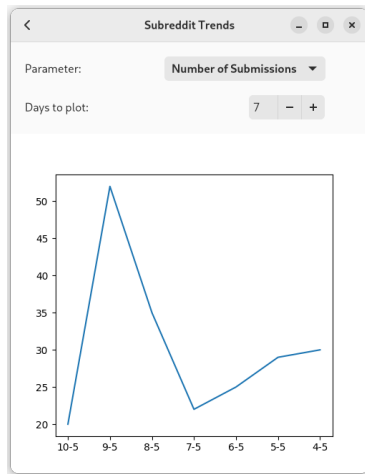
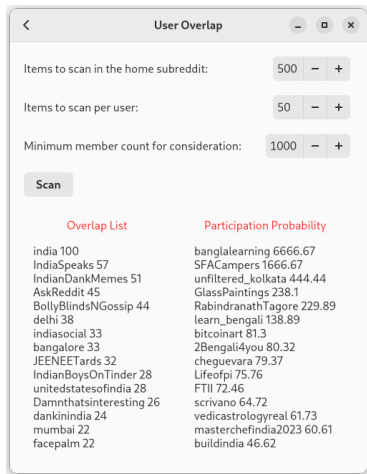
<sup>10</sup>[https://www.reddit.com/r/Flair\\_Helper/comments/ro0uqk/flair\\_helper\\_is\\_currently\\_not\\_accepting\\_new/](https://www.reddit.com/r/Flair_Helper/comments/ro0uqk/flair_helper_is_currently_not_accepting_new/)

# Objectives

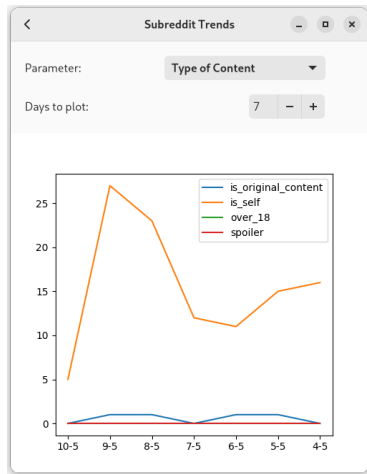
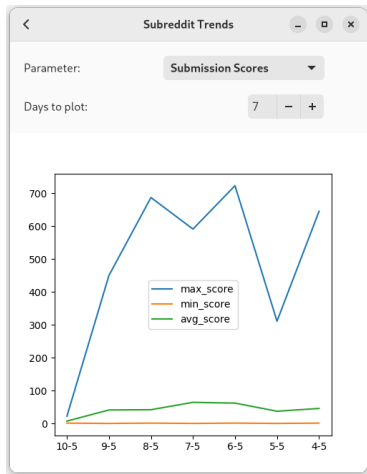
- A robust set of tools to deal with adverse scenarios
- Ability to customize the parameters with ease
- Access to logs to keep everything in check
- Meet performance requirements in general



## Data Collection



# Data Collection



# Bulk Action

<

Prune Members

- □ ×

Minimum Link Karma:

500

- +

Minimum Comment Karma:

100

- +

Minimum Account Age:

150

- +

Subreddit Blacklist:

Word Blacklist:

subreddit1  
subreddit2  
subreddit3  
subreddit4

word1  
word2  
phrase1  
phrase2

Prune

<

Purge Content

- □ ×

Posted in the last () days:

3

- +

Posted by banned users:

☒

Word Blacklist:

Domain Blacklist:

word1  
word2

domain1  
domain2

Purge

## Future Scope

- Make primary moderation features in place
- Run in background on a scheduling basis
- Integration with AutoModerator
- Integration with Pushshift dataset
- Train a model to suggest actions

# References

1. <https://www.lti.cs.cmu.edu/sites/default/files/shen%2C%20qinlan%20-%20Thesis.pdf>
2. <https://scholar.colorado.edu/downloads/4m90dv84k>
3. [http://www.eshwarchandrasekharan.com/uploads/3/8/0/4/38043045/eshwar\\_thesis.pdf](http://www.eshwarchandrasekharan.com/uploads/3/8/0/4/38043045/eshwar_thesis.pdf)
4. <https://smartech.gatech.edu/bitstream/handle/1853/62779/JHAVER-DISSERTATION-2020.pdf>
5. <https://ojs.aaai.org/index.php/ICWSM/article/download/3217/3085/6266>

# Thank You!