A case study on r/Malaysia

# Modernizing Community Moderation

Derek Chong

---

*Note: In place of a YouTube video I elected to demonstrate my results in the form of an interactive online demo that readers can try for themselves. Please refer to Classifying Promotion: Interactive Demo below for more information.*

## Introduction

> *"When the variety or complexity of the environment exceeds the capacity of a system (natural or artificial) the environment will dominate and ultimately destroy that system."*
> -- *Ross Ashby*

Ashby's Law of Requisite Variety, also known as the First Law of Cybernetics, requires that control systems be more complex and flexible than the system being controlled.

It is clear that in recent years, online communities have struggled to keep pace with the demands of environmental complexity. Online communities and discussion boards have become an essential part of the modern public commons, and are vital to the exchange of ideas and shared thought. But at the same time, every major platform seems to be failing to fulfil this ideal at scale, falling prey to political influence campaigns, manipulation for financial gain, filtering out reality, polarizing users, enabling trolling, and as an end result, damaging users' well-being.

A key component of a healthy online community is quality moderation, which frequently relies on the cognitive surplus of volunteer moderation teams. As this is only available in limited quantities, tooling and automation is essential to ensure that moderator efforts can be applied to maximally develop their communities. However, in practice tooling is surprisingly primitive. The state of the art appears to be the use of generically-tuned Bayesian spam filters and at most, rules-based engines, with growing interest and more advanced technologies used in some inhouse moderation teams:

> *"There are some areas where technical tools are helping us do this job […] but the vast majority, when we're looking at hate speech or we're looking at bullying or we're looking at harassment, there is a person looking at it and trying to determine what's*

*happening in that offline world and how that manifests itself online." – Monika Bickert, Facebook*

For my contest entry, I would like to examine the application of tools acquired in CS109 to improving the effectiveness of online community moderation, as well as to better understanding the underlying dynamics of the community userbase, particularly in the area of problematic or "toxic" users. I will do so in the context of a country-specific community on Reddit, of which I am a moderator, r/Malaysia.

# Background

## On r/Malaysia

I have been a moderator of Reddit's r/Malaysia, a country subreddit with about 80,000 users, for about two years. This has provided exposure to the inside perspective on the challenges of community-building and moderation as a community scales.

I intend to use the subreddit as a petri dish, in which we can run analysis and experiments that can later be scaled as desired. The subreddit will be helpful to this end in that while relatively small (receiving three million views a month), it is a fully-functioning community with a distinct culture and norms, where users operate in the same timezone. Metadata on users, content and moderation is readily available, and policies are much more comprehensive that across the site as a whole. Also importantly, I have buy-in from the moderation team to run analysis and experiments for the purposes of this project.

Access to comprehensive moderator data and context will be a key advantage in our analysis. Studies on social media studies are frequently conducted from an external sociological perspective, whereas we have access and understanding of every content removal, warning, ban, user annotation, user report and moderator discussion required. From my review of the academic literature, there seem to be relatively few studies which take advantage of community moderation data.

## User Toxicity

Online communities have learned over the years that moderation is essential in preventing the topics that generate the strongest emotions from taking over. Said topics tend to attract obsessive users, who hold strong, fixed opinions, and also have much more disposable time than normal users. Allowing this process to run unchecked eventually drives out regular users, which drives out regular

discussion, and sends communities into death spirals. Other kinds of misbehavior affects the quality of communities as well, through Broken Windows Theory - self-promotion, spam, hate speech, or things as simple as violating basic norms such as post format requirements (the equivalent of littering).

For the purposes of this report I will refer to such users as “toxic users”.

# Data and Cleansing

## Plan and Scope

We identified a data collection period which would provide data that is consistent and representative of the current state of the community: major subreddit moderation and rule changes were introduced in 2018 to combat toxicity as the subreddit grew in size (as subreddit subscriber count doubles every year), which were announced on 9 October 2018.

We collected data from this date to 25 May 2020, one week before the beginning of the data collection period, to ensure post and content scores settled to their final values, a duration of about 19 months.

We identified key sources of external data, which were both required in order to ensure comprehensive coverage:

- Official Reddit API: Provides live data, but limits results to 1,000 entries of any kind. No access to deleted posts, which are important for our purposes. Internal moderation data is available programmatically. However this API is severely rate limited (30 requests per minute), such that it is not viable to collect all data via this channel.
- Pushshift API: A third-party source which provides an extensive repository of reddit data at high throughput, but without accurate post and comment karma scores. Deleted posts available. Very occasionally missing a post or comment.
- YouTube API: Provides data at a relatively low rate limit per day.

## Data Collection

We collected data in bulk upfront, as collection lead time was a concern due to rate limits on external API: if some data was required later across a large range of users, we might not be able to retrieve it in a timely fashion.

We designed strategies for intelligent data collection. At 30 requests per minute, the 750,000 comments on r/Malaysia during the measurement period would take 416 hours to collect. We found ways to squeeze multiple requests into single calls, or maximizing the value of data returned per call (for example, making post requests and capturing the metadata for every comment provided, or squeezing 50 video IDs onto each YouTube API call).

We identified following sets of data as potentially useful for our analysis, and collected it across a dozen separate Python scripts and queries, taking over 2 days of runtime:

| Type | Item | Volumes | |
|---|---|---|---|
| Subreddit data | All posts | 44,487 posts, 213.9MB PSAW / 1.74GB PRAW | |
| | All comments | 774,713 comments, 1.68GB PSAW | |
| Moderation data | Ban list | 15.6MB Reddit API | |
| | Modmail | | |
| | Moderation log | | |
| | User annotations | | |
| User data | Metadata for all r/Malaysia users | 26,158 users, 40.5MB PRAW | |
| | Posts by users to any subreddit | 12.21GB PSAW | |
| | Comments by users to any subreddit | 31.49GB PSAW | |
| Youtube data | Video metadata for videos posted to sub | 2,125 videos | 5.8MB |
| | Channel metadata | 1,191 channels | 1.7MB |

# Data Cleansing

We found that it was extremely important to remove bots from dataset when running analyses. Bots are much more active than normal users and operate in irregular patterns and across timezones. We ran a multiple-pass approach:

- Picked out very well known bots
- Sorted comment by size and removed obvious bot names
- Sorted user activity by frequency and picked out obvious bot usernames
- Added list of known bots
- Sorted users by activity variance (see below)

We identified and removed around 250 bots.

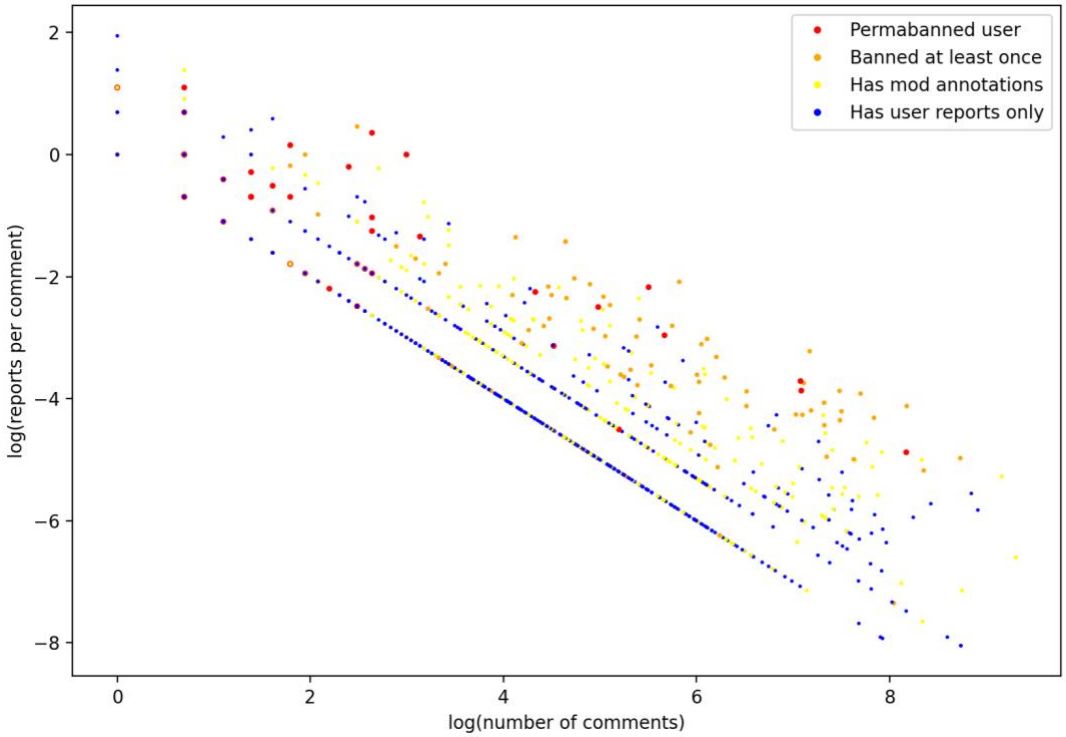
Where applicable, we ran passes to exclude where would affect particular questions, for example karma counts are not available for deleted Reddit comments, and some YouTube accounts hide their view and/or subscriber counts. Such removals are highlighted within subsections below.

# Scoring User Toxicity

In order to perform analysis on toxic users, we first need to have a metric for toxicity.

After considering methods of scoring based on quantifying the moderator interventions required per user (warning, content removals, bans, etc.), we discovered a simple and effective metric in the form of user reports.

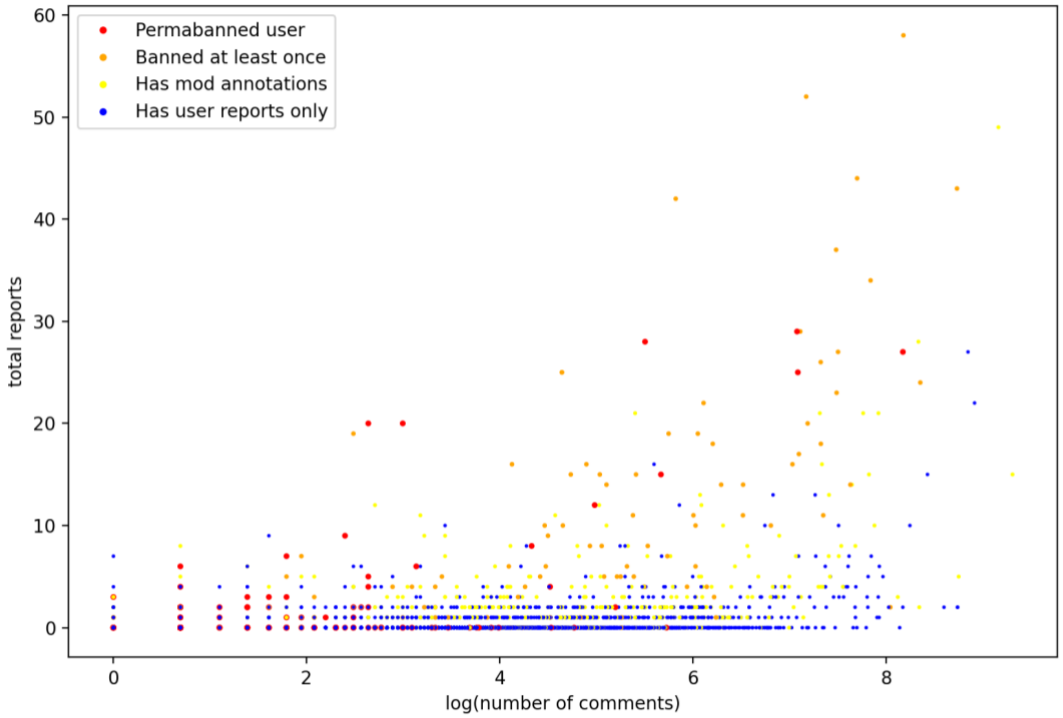
User reports are a feature that Reddit provides which enables users to submit anonymous reports on content that breaks subreddit or sitewide rules. We parsed all user reports from all comments made during the measurement period, and aggregated them for each user. We found that the average number of reports per comment that a user receives correlates quite well with moderator interventions:




I believe this is because the users that are aware of the reporting feature and use it regularly tend to be (a) more conscientious than average and (b) regulars of the subreddit community, who have been properly educated on community rules and norms. From a statistical point of view, users do not see who else has submitted a report, which means each report can be thought of as independent observations from the underlying distribution of the offending user's toxicity level. As such, although moderators frequently encounter inaccurate reports, at the aggregate level the Central Limit

Theorem may take loose effect on the sum of the reports on each user (despite the sample size being small for each user), causing a wisdom-of-crowds effect to emerge.

Note that the scatter plot above is placed on a log-log scale in base 10, as user interaction appears to follow a power law, where a small minority of users contribute an exponential volume of content. A log-linear plot of absolute reports for comparison:




To further verify this finding, we examine expected report ratios conditioned on moderator interventions:

$$E[\text{reports per comment} \mid \text{user permabanned}] = 0.151$$
$$E[\text{reports per comment} \mid \text{user banned at least once}] = 0.135$$
$$E[\text{reports per comment} \mid \text{user has mod annotations}] = 0.0807$$
$$E[\text{reports per comment} \mid \text{user has user reports only}] = 0.00590$$

Additional insights from interpreting the data and plots above:

- As the number of reports per comment increases, users become increasingly likely to encounter mod interventions. Setting aside users with zero reports and/or zero comments, the correlation coefficient on the first plot above is -0.91, which is very strong.
- User accounts exist in a surprisingly tight band of reports per comment. If a user is constantly receiving reports they tend to be removed from the community. This suggests that moderation decisions are generally fair.

- However, a certain amount of tolerance for reports seems to emerge as a user account becomes more prolific. This might be due to longtime users being better at gaming the limits of acceptable behavior, or increased moderator tolerance for the misbehavior of regular users.

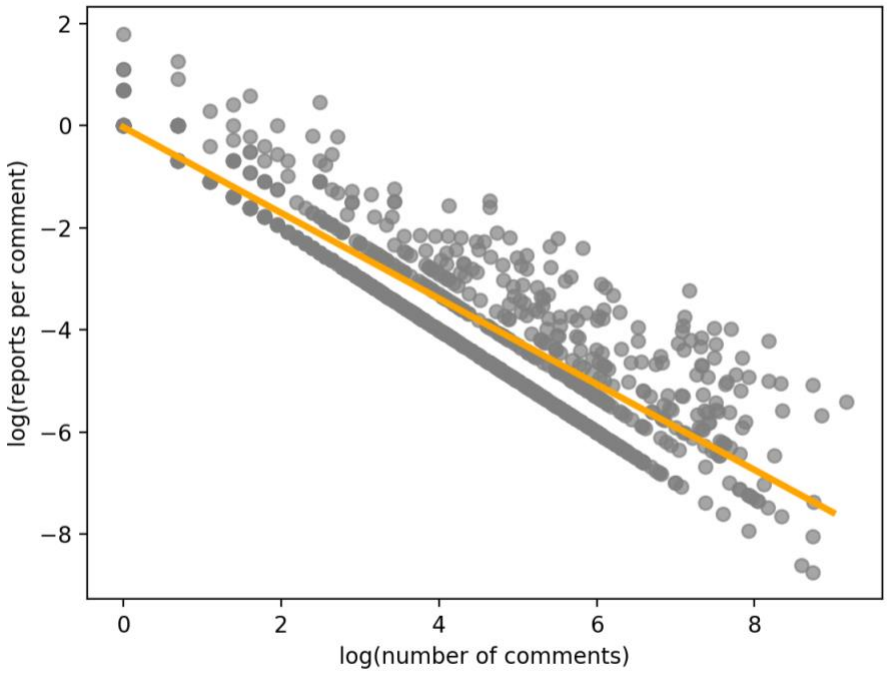
- Toxic users delete their accounts over time: We found a significant number of comments from deleted accounts (where the username was not retrievable), where the average reports per comment was an extreme outlier, and removed from the above plots.

- Most users are "good": Out of the 23,832 users who commented on the subreddit during the measurement period, only 1,019 received any reports at all!

Based on the above, we conclude that although more finely-tuned scoring metrics may be developed with additional study, user reports may be used as a reasonable basis for measuring or classifying toxicity in our following analyses.

For classification purposes, we ran a linear regression on the logs of the data above, which returned the following:

*log(predicted reports per comment) = -0.0342*log(total comments) - 0.838*




We classified users as toxic if they scored greater than one point higher than predicted by the above equation given above. We then examined the resulting set of users identified and verified that this cutoff was reasonable: most of the users returned were generally known to be problematic by the

moderation team. This sanity check also identified several toxic users which were not yet known to the team.

Given the success of this approach and time constraints, we elected to use this set of users for our following analyses without applying a more rigorous cutoff (e.g.: using standard deviation to set the classification cutoff at a particular percentile of the userbase).

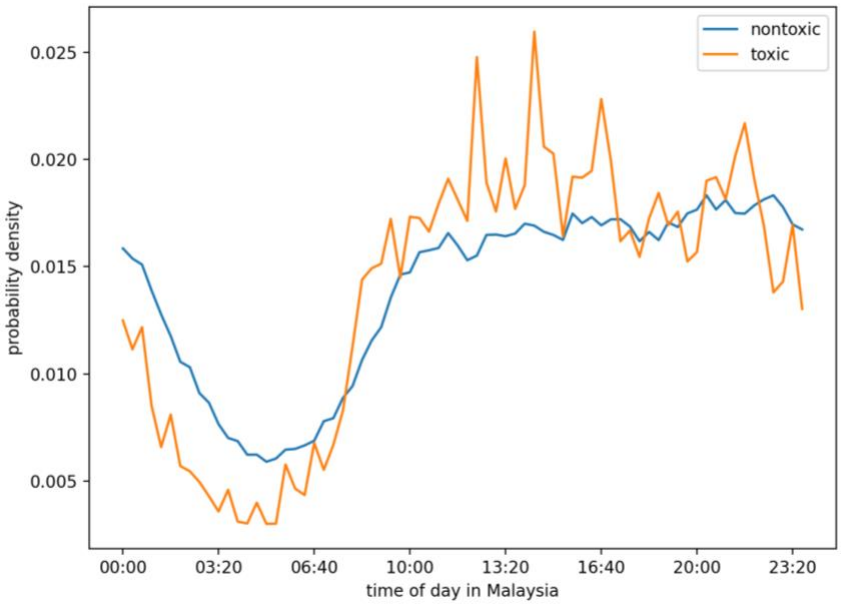
# Toxicity and Sleep

It is extremely well-established in the medical literature that poor sleep is associated with mental health issues of all kinds, including depression, anxiety, mood disorders, OCD, ADHD and schizophrenia. Sleep disorders have been described as a core symptom of depression, wherein three quarters of depressed patients have symptoms of insomnia. Mood disorders are found in one third to half of patients with chronic sleep problems. Causally, chronic insomnia is a risk factor for developing anxiety disorders and depression, and sleep interventions reduce the severity of mental health symptoms.

Given the above, one question that we might ask our dataset is whether toxic users sleep more poorly than regular users. Users accumulate hundreds to thousands of timestamped posts and comments per year in their engagement with the subreddit, which when aggregated, can be used to shed light on their patterns of sleep at both the individual and aggregate levels. Finding poorer patterns of sleep might suggest that one root cause of toxic behaviour online is the existence of underlying mental health issues.

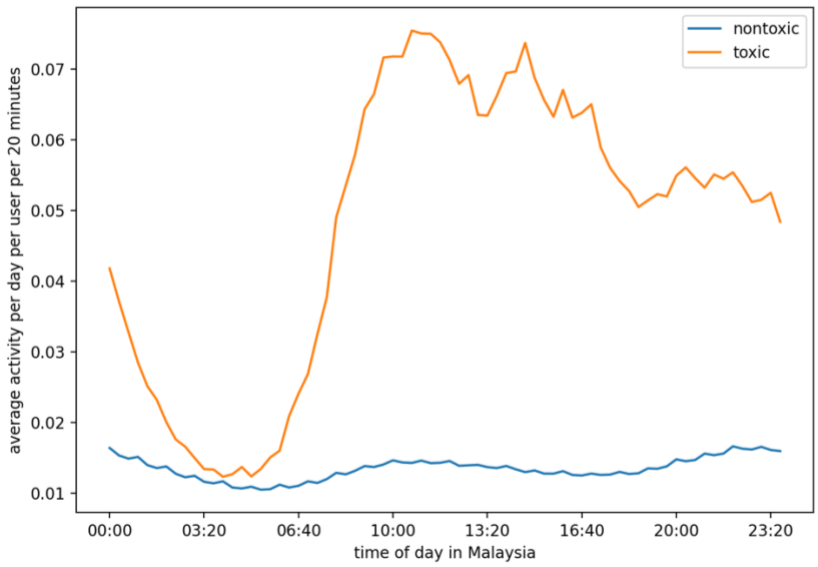
# Aggregate Sleep Statistics

We classified users into toxic and non-toxic groups as previously described, and plotted the distribution of both groups' activity by time of day in Malaysia, across all activity in each user's Reddit account. We normalised activity levels such that each user contributed exactly *1/n* to their plot, and accumulated activity into 20-minute buckets to improve comprehensibility:



We found surprisingly clear differences between the shape of the activity patterns of the toxic and nontoxic users. Toxic users go to bed and wake up one hour earlier than non-toxic users, and appear to have a slightly wider spread in their windows of sleep. We observed experimentally that this difference increased when we set a higher toxicity classification cutoff, suggesting that this is a true feature of toxic users.

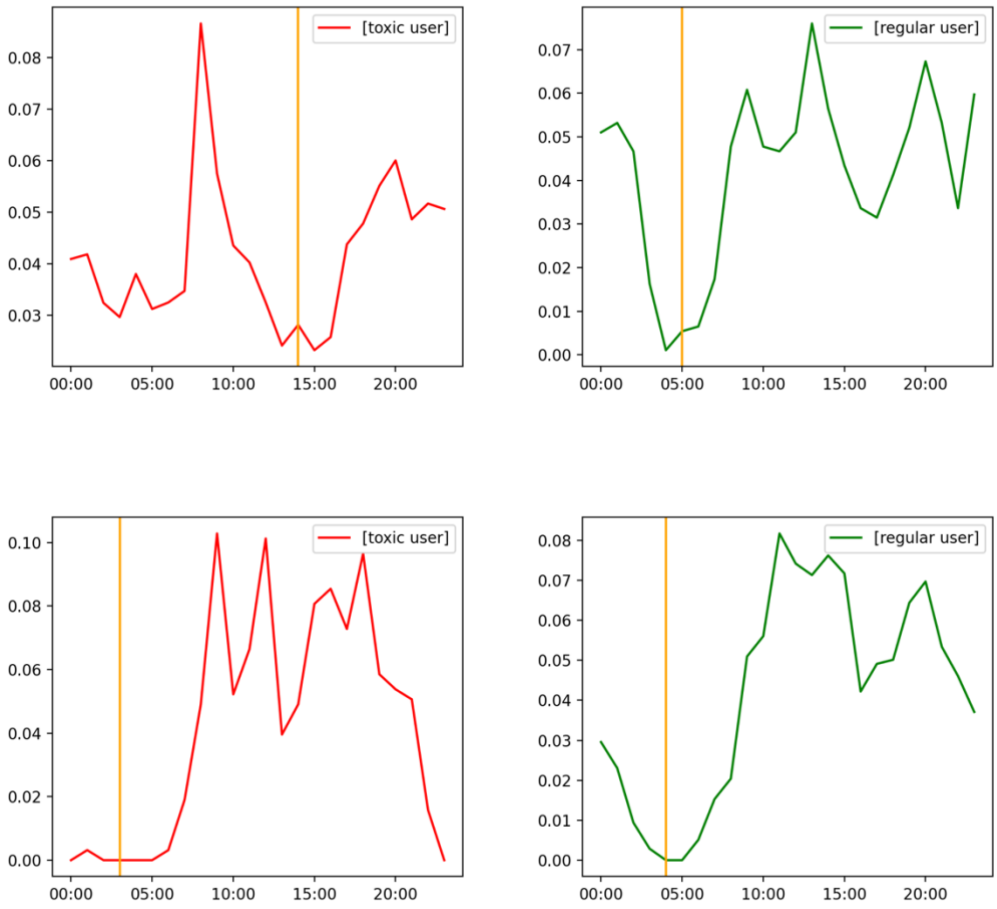
The diagram above seems to show that toxic users participate less than regular users during the night (and more during the day). We confirm that this is an artifact of the normalization process by plotting in terms of activity per user, which reveals that the previous is actually caused by the dramatically higher activity of toxic users during the daytime as compared to regular users:



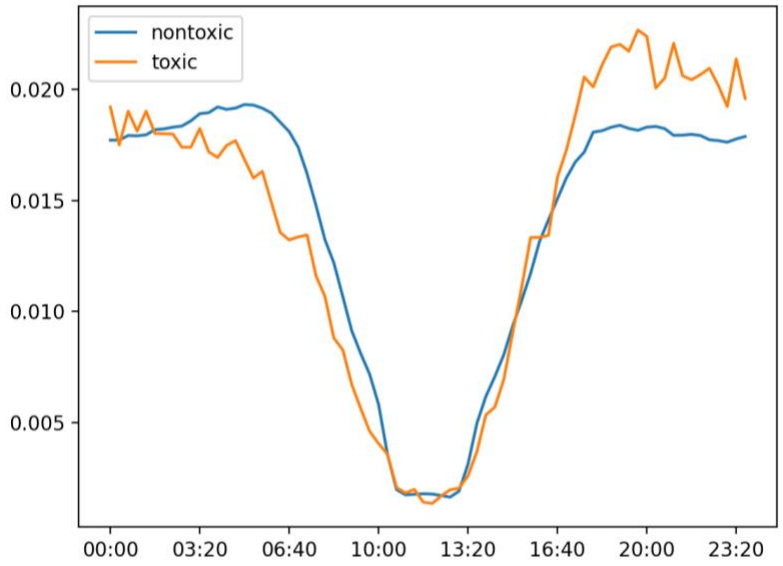

To further examine the differences in width of sleep windows, we developed a technique for normalising sleep timing across users, regardless of each user's actual timezone or working hours:

- We restrict our dataset to users with at least 50 activities within the measurement period, to ensure there is a reasonable amount of sample data.
- For each user, we bucket their activities individually into one-hour buckets across their day.
- Given that sleep is a physiological need, users *must* sleep for at least a contiguous four-hour period somewhere during their day. We identify the period in which the sum of activity for three hours is on average the lowest, and label the centre of this period the user's peak sleep time.

We found that this process produced quite good results at the level of individual users (charts provided with usernames omitted for privacy reasons):




Once we have all users' peak sleep times, we shift each users' activity timestamps such that their peak sleep time appears at noon, and reaggregate the data for both sets of users:



This plot confirms that even once possible differences in time zones and working hours are accounted for, on toxic users have a wider variance of activity around their peak sleep hours (an additional range of 45 to 120 minutes), suggesting poorer sleep hygiene.

## Individual Sleep Statistics

Now that we can identify the peak sleep time of individual users, we can score sleep quality on a user-by-user basis.

For each user, we measure the amount of activity per night that the user engages in which occurs in a seven-hour window centred around their peak sleep time. The National Sleep Foundation suggests that adults require a minimum of seven hours of sleep per night. If a user is regularly active on Reddit during the window in which they are most frequently sleeping, this implies that their sleep is disturbed in some way, such as generally poor sleep hygiene, or medical conditions such as insomnia, frequent night waking, or delayed phase syndrome.

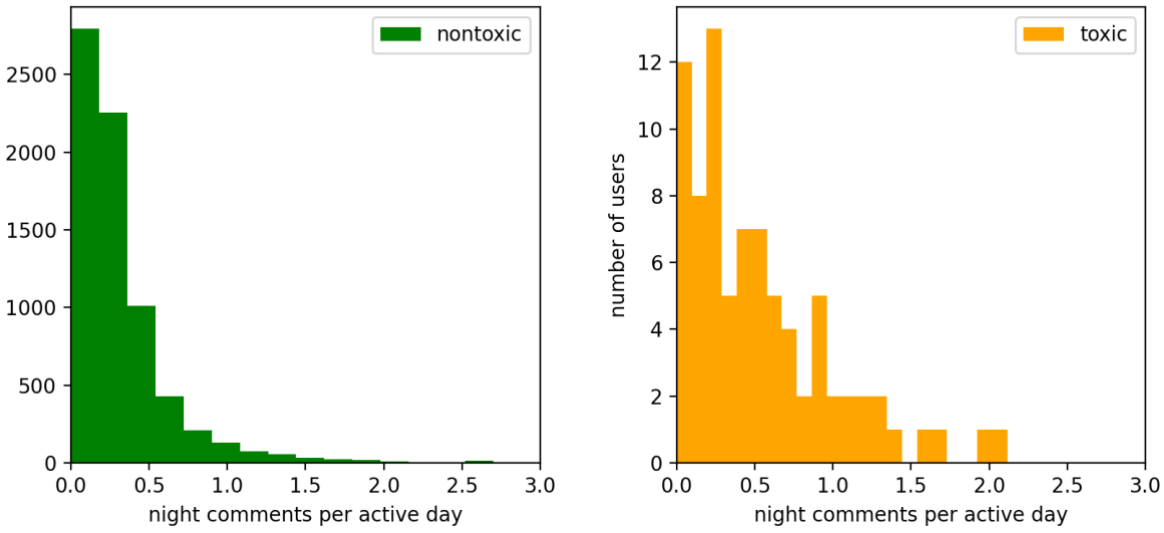
For the purposes of this exercise we exclude both users who have less than 50 activity datapoints as above (we cannot be confident about their average sleep time), and users who have less than 5 activity datapoints specifically on our subreddit, because while they may exhibit sleep disturbances, we are unlikely to have enough experience with this user on our subreddit to classify them as toxic, even if they are extremely toxic across other subreddits.

We find the following statistics for the different sets of users:

| | **Toxic** | **Non-Toxic** | **Combined** |
|---|---|---|---|
| **# Users** | 83 | 7,129 | 7,212 |
| **Nights active per days active** | 0.228 | 0.172 | 0.173 |
| **Night comments per days active** | 0.618 | 0.354 | 0.357 |
| **Night comments per nights active** | 2.22 | 1.70 | 1.70 |

These results show that toxic users are dramatically more active at night than non-toxic users, per day that they are active on Reddit as a whole. Toxic users are active on Reddit during their regular seven-hour sleep window 33% more often than non-toxic users, and on the nights that they are up, they make 31% more comments, resulting in a total of 75% more night comments per day

Plotting the histograms of each set of users' night comments per day, we see that the distributions look somewhat distinct:




We used the bootstrap method to estimate p-values for the probability that both of these samples were drawn from the same underlying population. We combined the data from both samples into a single universal distribution, drew samples of size 83 from it 10,000 times at random, computed the sample mean for each sample, and recorded the number of times this was greater than the mean of the observed toxic users sample.

The resulting p-value was 0.0005, which suggests that the sleep statistics of the toxic users sample was very unlikely to have arisen by chance.

We considered that this effect might be occurring due to the fact that toxic users tend to be more active on Reddit than the average user, and confirmed that the effect persisted and continued to be statistically significant when we calculated bootstrap p-values against a sample of nontoxic users randomly selected to have a similar sample mean, with $p = 0.0034$.

# Classifying Self-Promotion

Managing self-promotional activity is another area where I believe that tools learned in CS109 might be applied to improve the state of online communities. Moderators regularly spend time removing self-promotional content on communities across the internet. This content hurts the signal-to-noise ratio of the community and is not always easy to identify using a generic spam filter, and self-promoters have strong incentives to find ways to post their content.

My experience of the decision process for this activity is that identifying self-promotion correctly requires some intelligence, but essentially boils down to considering the values of a few inputs. This seems like a task that might be tractable for a probabilistic classifier.

One major block of self-promotion removals is users posting their own content from YouTube. Querying our dataset revealed that 2,375 of the 44,487 submissions in our measurement period were YouTube posts, and a surprisingly high 1,059 of these 2,375 posts were later removed by moderators. As this seems like a healthy amount of data for training purposes, we decided to constrain the scope of this exercise to self-promotion via YouTube (although other classes of self-promotion can be handled similarly).

## Features and Data Collection

To begin, we considered the information we personally use when making decisions about whether a piece of content is self-promotion, and how best to encode these in a numeric format. We identified the following features that might be easily retrieved in real-time by a classifier script:

| | | |
|---|---|---|
| 1 | Name similarity | How similar is the name of the user to the YouTube channel? |
| 2 | Subscriber count | How many subscribers does the YouTube channel have? |
| 3 | View count | How many views does the YouTube video have? |
| 4 | YouTube post ratio | What percentage of the user's recent posts are from YouTube? |
| 5 | Reddit karma | How much karma does the user's account have? |
| 6 | Account vs. post date | How long has this Reddit account existed for prior to this post? |
| 7 | Post vs. publish date | How long ago was the YouTube video published relative to when it was posted? |
| 8 | YouTube channel | Does this YouTube channel have a track record of non-self-promotional content? |

We next wrote Python scripts to collect values for each feature, for every post in our training set. We set aside posts where any item was not available, for example in the cases where users deleted their accounts, where the YouTube channel was set to hide its statistics, or where the post was actually a YouTube channel, user or playlist instead of a video.

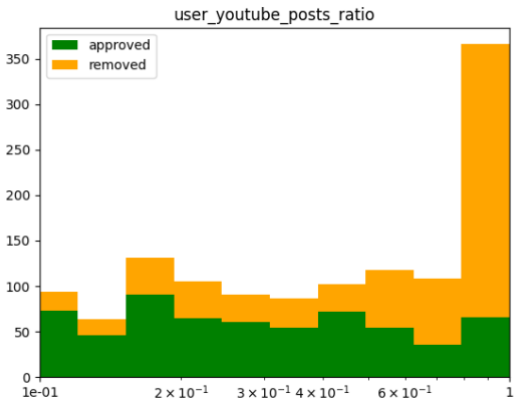
We also collected the approval/rejection decision made by the moderation team and made the choice to use this as the sample classification for training purposes. While not all video removals are for self-promotional reasons, this is an authoritative source of truth, and may encode other information that the classifier might be able to make use of (also, hand-classifying 2,000 videos is nontrivial!).

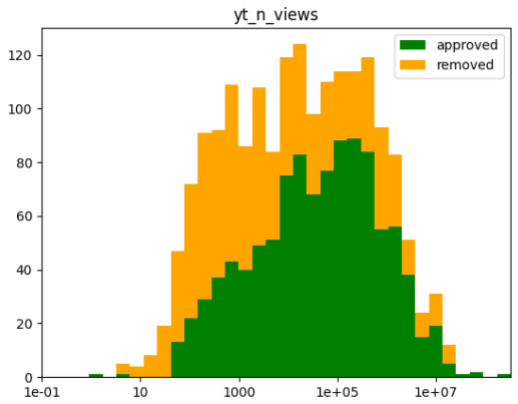
After all collection and validation, the resulting dataset contained 1,823 rows of post data. We set aside a random 20% of this data to serve as test data for validating classifier performance.

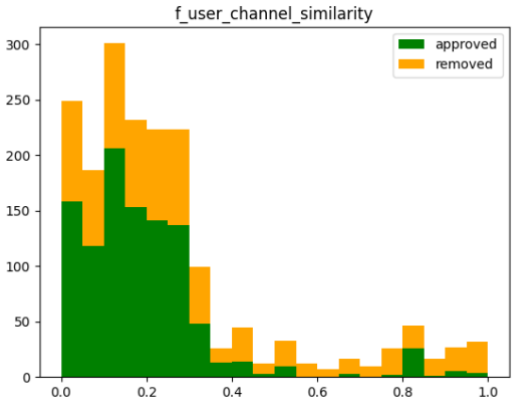
| id | url | author | title | created | yt_channel_i | yt_channel_i | yt_title | mod_remove | yt_attributec | yt_n_views | yt_pub_date | yt_channel_i | yt_channel_i | yt_channel_r | user_link_ka | user_comm | user_creatio | user_n_posts | user_n_posts | user_n_youti | user_youtub | exclude_test | f_user_chani | f_diff_use |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| gpw36y | https://youti | arquish | Kami Dari Dc | 1590349784 | Rave Ryders | https://www | Sesi ucapan raya dari perantauan 2020 | | | 65 | 1590173467 | 0 | 40 | Rave Ryders | 164 | 1 | 1529354266 | 5 | 1 | 2 | 0.4 | N | 0.33333333 | 608192 |
| gpr3ss | https://www | kwokdexter | Old footage | 1590332557 | City Rails | https://www | Touch' n Go [ LRT Kelana Jaya ] QR Cod | | | 323 | 1590289222 | 1250 | 174 | City Rails | 12615 | 3437 | 1529468654 | 283 | 1 | 11 | 0.03886926 | N | 0.1 | 608205 |
| gpqixe | https://youti | arquish | Salam peran | 1590330556 | Nazmi Zahri | https://www | Raya Di Pera | Y | | 9 | 1590293739 | 0 | 3 | Nazmi Zahri | 164 | 1 | 1529354266 | 5 | 1 | 2 | 0.4 | N | 0.33333333 | 609394 |
| gpncp6 | https://youti | highlands2h | $20 A NIGHT | 1590315525 | highlands2h | https://www | $20 A NIGHT | Y | | 1212 | 1588912421 | 392 | 80 | highlands2h | 56 | 26 | 1574195586 | 7 | 2 | 5 | 0.71428571 | N | 1 | 127168 |
| gpktho | https://www | DuoFusionz | I dont know i | 1590301498 | | | | Y | | 173061 | 1590018411 | 36300 | 100 | RMP TV Chai | 4 | 0 | 1553477888 | 2 | 1 | 1 | 0.5 | N | 0.08333333 | 365405 |
| gpkkb7 | https://youti | CastleCarv | Wrote a vide | 1590300171 | Amin Farrkh | https://www | So I've been | Y | | 46 | 1590232553 | 9 | 1 | Amin Farrkh | 259 | 1992 | 1372675178 | 17 | 1 | 1 | 0.05882353 | Y | 0.17391304 | 2175573 |
| gpkjtq | https://www | aimanrijal | Selamat Har | 1590300101 | Space Recon | https://www | Alwagera - S Y | | | 479 | 1590199232 | 449 | 5 | Space Recon | 1 | 0 | 1545749769 | 6 | 4 | 5 | 0.83333333 | Y | 0.15384615 | 444494 |
| gpvf4 | https://youti | anondnanym | The most wi | 1590292201 | Zafran Rarar | https://www | Suasana Virtual Hari Raya - #MuzikDari | | | 455 | 1590187870 | 361 | 37 | Zafran Rarar | 241 | 64 | 1552040048 | 29 | 1 | 12 | 0.4137931 | Y | 0.28571429 | 381478 |
| gphruh | https://youti | CaptainzNim | Selamat hari | 1590287332 | Captain Nim | https://www | Ku Pohon Restu Ayah Bonda Cover Di Pi | | | 72 | 1590250020 | 271 | 39 | Mohd Naim | 656 | 60 | 1562789844 | 54 | 1 | 16 | 0.2962963 | N | 0.26086957 | 274601 |
| gpgfpb | https://www | ArmedSweej | 1 HOUR OF I | 1590281793 | MothershipS | https://www | 1 HOUR OF LOFI RAYA BEATS TO LISTE | | | 68600 | 1590194713 | 43000 | 758 | MothershipS | 12878 | 1830 | 1512894586 | 181 | 1 | 3 | 0.01657459 | N | 0.16 | 773001 |
| gpfwnw | https://www | gustavowing | South China | 1590279757 | Shaoyu Yuan | https://www | South China | Y | | 1086 | 1589965784 | 330 | 11 | Shaoyu Yuan | 182 | 2 | 1495923150 | 37 | 14 | 35 | 0.94594595 | N | 0.27272727 | 940426 |
| gpbs17 | https://youti | haz__man | Everyone knc | 1590254310 | Nadya Said | https://www | P Ramlee - Dendang Perantau (LIRIK LA | | | 26498 | 1528826580 | 24800 | 24 | Nadya Said | 11048 | 23350 | 1436696287 | 261 | 1 | 13 | 0.04980843 | Y | 0.22222222 | 921302 |
| gp8lyl | https://www | tha_oz | Selamat Har | 1590254112 | JimyboyTV | https://www | Our First Time | | | 125173 | 1590196920 | 1080000 | 235 | JimyboyTV | 915 | 532 | 1552978707 | 117 | 1 | 2 | 0.01709402 | N | 0.125 | 372182 |
| gp7guw | https://www | shodaimezac | selamat hari | 1590250033 | RMP TV Chai | https://www | SYAWAL YANG BIRU - POLIS DIRAJA M | | | 173061 | 1590018411 | 36300 | 100 | RMP TV Chai | 2048 | 4759 | 1554195301 | 110 | 1 | 14 | 0.12727273 | N | 0.23076923 | 358231 |
| gp1h87 | https://youti | BDTchannell | Wanita meg | 1590222631 | | | | Y | | 137 | 1590087375 | 0 | 204 | BDT Channel | 1 | 0 | 1589993128 | 4 | 2 | 4 | 1 | Y | 0.7826087 | 942 |
| gp05ai | https://youti | KedaulatanS | Speak Malay | 1590215146 | Sierra Lisse | https://www | Speak Malay Like a Local - Lesson 78 : \ | | | 1764 | 1589908704 | 12800 | 96 | Sierra Lisse | 499 | 624 | 1583332205 | 9 | 2 | 3 | 0.33333333 | N | 0.15384615 | 65764 |
| govzpa | https://www | AlexVostox | [OFFICIAL FR | 1590196583 | Metrowealth | https://www | KONGSI FULL MOVIE | | | 2081055 | 1589684411 | 309000 | 2519 | Metrowealth | 19957 | 23638 | 1569212938 | 204 | 1 | 32 | 0.15686275 | N | 0.20689655 | 204714 |
| gotegv | https://youti | AirMonkey11 | Malaysian tr | 1590186917 | AP Archive | https://www | SYND 18 6 76 JOINT THAI / MALAYSIAN | | | 803 | 1437707559 | 2710000 | 484098 | AP Archive | 3361 | 1974 | 1575185721 | 84 | 1 | 10 | 0.11904762 | N | 0.26086957 | 1374781 |
| gos4wi | https://youti | welshwank | Our national | 1590182579 | Basil Henriq | https://www | Mamula Moon-Terang Bulan-Terang bc | | | 107546 | 1328158490 | 1200 | 127 | Basil Henriq | 310 | 783 | 1483082828 | 10 | 1 | 1 | 0.1 | N | 0.16666667 | 1549243 |
| gopy5a | https://www | imperialx87 | Yo guys chec | 1590175428 | Anodyne Atn | https://www | Kimi, Rapiid | Y | | 677 | 1590117478 | 216 | 6 | Anodyne Atn | 17 | 2 | 1526214696 | 7 | 2 | 2 | 0.28571429 | N | 0.20689655 | 639027 |
| goiaau | https://youti | Kirito_0094 | Raya annouc | 1590153858 | Buletin TV3 | https://www | Aidilfitri Pada 24 Mei 2020 | | | 24334 | 1590113058 | 683000 | 30680 | Buletin TV3 | 127 | 381 | 1569791691 | 72 | 1 | 8 | 0.11111111 | N | 0.09090909 | 203213 |
| gogns1 | https://youti | lelarentaka | College hum | 1590142322 | CollegeHumc | https://www | How Did YOU Y | | | 191309 | 1590046201 | 13800000 | 3607 | CollegeHumc | 9023 | 172964 | 1368818704 | 185 | 1 | 20 | 0.10810811 | N | 0.26086957 | 2212274 |
| gogica | https://youti | cindyhoo | I tried to ma | 1590141580 | Malaysian Ki | https://www | ◆üáś◆üáń◆Y | | | 142 | 1590041033 | 34 | 4 | Malaysian Ki | 1 | 0 | 1589509026 | 3 | 2 | 3 | 1 | Y | 0.19354839 | 5320 |
| godow1 | https://www | Hairy_Tes_te | GERAK PANT | 1590126999 | Soldier From | https://www | GERAK PANTAS 2018 - MALAYSIA PARA | | | 92988 | 1536086689 | 1840 | 4 | Soldier From | 125 | 115 | 1589509986 | 9 | 1 | 3 | 0.33333333 | N | 0.15384615 | 534212 |
| gobkb8 | https://youti | kwokdexter | Lockdown: N | 1590117592 | VeggiesOnTc | https://www | Lockdown: Malaysia the safest country | | | 240597 | 1589862639 | 3780 | 23 | VeggiesOnTc | 12615 | 3437 | 1529468654 | 283 | 1 | 11 | 0.03886926 | N | 0.26086957 | 603939 |
| goi1ur | https://youti | AlexVostox | F/A-18D v.s | 1590083427 | The Ready R | https://www | F/A-18 vs MiG-29 - Dogfighting the Fu | | | 165805 | 1589970542 | 7850 | 9 | The Ready R | 19957 | 23638 | 1569212938 | 204 | 1 | 32 | 0.15686275 | N | 0.25 | 207576 |
| gnwiew | https://youti | StatStorm | Southeast As | 1590066716 | StatStorm | https://www | Southeast As | Y | | 401 | 1589938200 | 192 | 161 | StatStorm | 46 | 38 | 1569083136 | 41 | 8 | 36 | 0.87804878 | N | 1 | 208550 |
| gnsr55 | https://www | dabongsa | Highway traf | 1590048449 | | | | Y | | 19957 | 1590011611 | 769000 | 2745 | The Star | 4359 | 25936 | 1491149941 | 205 | 1 | 11 | 0.05365854 | N | 0.125 | 988616 |
| gnrsqw | https://youti | Khidirsanshi | Raya tak lam | 1590043286 | Khidir San | https://www | DRAMA ANIF Y | | | 1262 | 1589975753 | 2610 | 24 | Khidir San | 2 | 4 | 1588313085 | 8 | 3 | 7 | 0.875 | N | 0.63636364 | 16626 |
| gnn3in | https://www | el_empty | KENDURI KA | 1590022784 | GTA 5 MALA | https://www | KENDURI RAHWIN TIME PKP \| BERBAL | | | 388929 | 1589605142 | 44200 | 27 | GTA 5 MALA | 2199 | 2736 | 1584223738 | 42 | 1 | 7 | 0.16666667 | N | 0 | 53814 |
| gnfgsr | https://youti | BDTchannell | Komik komik | 1589994971 | BDT Channel | https://www | S Adaptasi K Y | | | 77 | 1589759198 | 0 | 204 | BDT Channel | 1 | 0 | 1589993128 | 4 | 1 | 4 | 1 | Y | 0.7826087 | 2339 |
| gnb2vi | https://youti | Finnsravens | Raya anime | 1589980311 | | | | | | 4152644 | 1589942831 | 163000 | 733 | PETRONAS | 17 | 109 | 1529765486 | 1 | 1 | 1 | 1 | Y | 0 | 601773 |
| gn7vy0 | https://youti | kenwood-bn | Restui Kami | 1589964978 | BKTD Studio | https://www | VIDEO MUZIK AIDILFITRI 2020 BADAN I | | | 18177 | 1589858397 | 1160 | 18 | BKTD Studio | 694 | 1921 | 1586426521 | 13 | 1 | 1 | 0.07692308 | N | 0.06896552 | 34318 |
| gn6nhb | https://youti | rameshkulac | Eerie True G | 1589958594 | ramesh kula | https://www | A Dream Of Y | | | 603 | 1587112878 | 127 | 33 | ramesh kula | 1 | 0 | 1589205067 | 26 | 25 | 25 | 0.96153846 | N | 0.68965517 | 20921 |
| gn6h03 | https://youti | omnweift | Legend Of As | 1589956647 | All For One F | https://www | Legend Of Ancient Borneo - Animated | | | 22400 | 1589760295 | 810 | 87 | All For One F | 626 | 169 | 1490676685 | 65 | 1 | 2 | 0.03076923 | N | 0.19354839 | 990836 |
| gn5kia | https://youti | bananacranb | Heard their s | 1589953233 | | | | Y | | 26369 | 1544315077 | 362 | 12 | Shadow Pupp | 15 | 31 | 1569641957 | 6 | 4 | 4 | 0.66666667 | N | 0.12765957 | 253268 |
| gms8bi | https://youti | RossAnders | Cycladic Civil | 1589906792 | History and / | https://www | Cycladic Civil Y | | | 2965 | 1588676276 | 120 | 3 | History and / | 258 | 153 | 1577663231 | 85 | 21 | 69 | 0.81176471 | N | 0.32258065 | 110130 |
| gmreya | https://youti | matchaunag | Not from Mz | 1589904259 | Khong Guan | https://www | Sebuah Kena Y | | | 1038481 | 1589830391 | 31900 | 51 | Khong Guan | 39487 | 7254 | 1515429933 | 435 | 1 | 1 | 0.00229885 | N | 0.09090909 | 744004 |
| gmoo7n | https://youti | Shubotube | Do you know | 1589895320 | Shubotube | https://www | GRAPES \| Im Y | | | 364 | 1589851968 | 1790 | 103 | Shubotube | 918 | 17 | 1578841740 | 270 | 36 | 249 | 0.92222222 | Y | 1 | 130102 |
| gmnu8u | https://www | JJGabriel02 | I played Suas | 1589892254 | Jj Gabriel | https://www | Suasana Hari Y | | | 5871 | 1588654601 | 378 | 54 | Jj Gabriel | 20 | 3 | 1578729389 | 40 | 1 | 33 | 0.825 | N | 0.85714286 | 99252 |
| gmnnkl | https://www | JJGabriel02 | I played Suas | 1589887342 | Jj Gabriel | https://www | Suasana Hari Y | | | 5871 | 1588654601 | 378 | 54 | Jj Gabriel | 20 | 3 | 1578729389 | 40 | 2 | 33 | 0.825 | N | 0.85714286 | 99252 |
| gmkaix | https://m.yo | Wvrm_McFl | Another leak | 1589875216 | | | | | | 27522 | 1589835828 | 1720 | 127 | Tun Cardos | 20606 | 23963 | 1459735711 | 1060 | 0 | 26 | 0.0245283 | N | 0.2 | 1301001 |

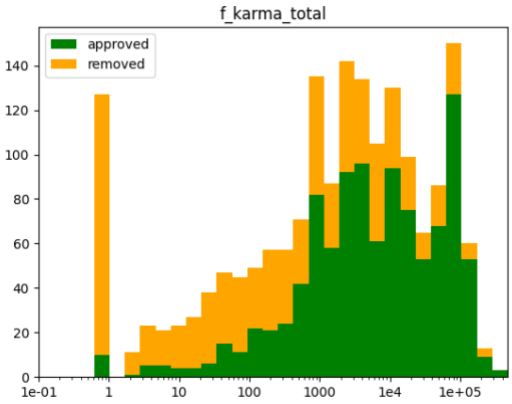
# Naïve Bayesian Classifier

As several of the features above are continuous in nature (ranging over a large set of values) with no clear parametric form, and Naïve Bayes classifiers prefer discretized data, we decided to create handmade buckets for each feature. We generated histograms for each feature, plotting all approvals and removals against feature values, for every row in our training set:






Note that plots are on a logarithmic scale, as much of the data was exponential in nature (such as YouTube view and subscriber counts).

We decided on 2-4 buckets for each feature, selecting bucket cutoffs by looking for clear delineations in the histogram data and applying domain knowledge. We then bucketed the training data accordingly and hand-implemented the Naïve Bayes classifier.

We ran the classifier on the test dataset and received the following results:

| | **Precision** | **Recall** |
|---|---|---|
| **Approvals** | 0.70383275 | 0.91402715 |
| **Removals** | 0.81553398 | 0.49704142 |

Interpreting the above: Our classifier does not appear to be not very accurate as a whole, but turns out to be quite suitable for the task at hand.

- It correctly approves 91.4% of all posts that would have been be approved, meaning minimal risk of widespread incorrect removals, which are more harmful than incorrect approvals, as they become hard for the moderation team to spot and correct.
- It correctly removes only 49.7% of all posts that would have been removed. This may be an outcome consistent with the fact that we trained the classifier on data that was not all self-promotional.
- 81.6% of all the posts it removes are actually self-promotion.

On further assessment we believe that the performance of the classifier may actually be closer to 95-99% in production:

- The classifier appears to have correctly understood the underlying self-promotion “signal”: When we looked at the supposed false positives: (a) more than half were items that should

have been removed but were missed by moderators, (b) many of the users were tagged as probable self-promoters by moderators despite allowing one video, (c) some of the posts were self-deleted, meaning moderators did not get to see them, and (d) the remaining items were exceptions that moderators made to the rules!

- Feature buckets in the training set were hand-configured with production in mind: For example YouTube videos from 18 months ago have view counts that are significantly higher than freshly-submitted self-promotional material.

Based on the assessment above, we made the decision to move forward with implementing and deploying the classifier bot on the subreddit.

# Other Classifiers

We also tried training a Logistic Regression classifier on our dataset, using scikit-learn instead of by hand (saving time, while learning a new software package). Following a similar training procedure, we received the following results on the test set:

| | **Precision** | **Recall** |
|---|---|---|
| **Approvals** | 0.73378840 | 0.83657588 |
| **Removals** | 0.74233129 | 0.60804020 |

Strikingly, despite this classifier being more suitable for continuous inputs, it performs similarly or worse on every dimension except for recall of Removals, meaning it identifies more items that should actually be removed.

Given that Logistic Regression works better with linearity in its inputs, we tried retraining the classifier on the dataset with logarithms applied to features with exponential components. This improved the results several by 2-5% on every evaluation criteria:

| | **Precision** | **Recall** |
|---|---|---|
| **Approvals** | 0.76206897 | 0.85992218 |
| **Removals** | 0.78313253 | 0.65326633 |

However, Approvals recall is a key dimension for our application and in this, the classifier still performs worse than a Naïve Bayes classifier.

Since scikit-learn requires input data to be standardized to a mean of 0 and standard deviation of 1, we can examine the weights of the classifier to learn about the importance of each feature. The features which affect the result the most are posting the same video repeatedly (0.69), rapid posting after account creation (0.51) and the YouTube channel being marked as trusted (0.39). Somewhat

surprisingly, the classifier gets almost no value out of knowing how long ago the video was published (0.08).

For interest, we finally also tried applying our data to scikit-learn's Random Forest classifier, which seems to be well-regarded and only required a few lines of code to plug in:

| | **Precision** | **Recall** |
|---|---|---|
| **Approvals** | 0.77777778 | 0.84435798 |
| **Removals** | 0.77401130 | 0.68844221 |

While this classifier outperforms the Logistic Regression classifier on most dimensions, and both classifiers seem to be identifying additional signal within the removals data that is lost in the process of hand-bucketing, such that they correctly classify an additional 18% of all removals, Naïve Bayes continues to perform the best on the key evaluation criteria for our task.

# Bot Deployment

We designed a bot as a serverless AWS Lambda function, to avoid the overhead of hosting and managing a server or virtual machine. The function is configured to automatically trigger every five minutes, and consumes under 1 GB-s of processing time per call, which means it can be run indefinitely on the AWS free tier.

We implemented the orchestration, processing and data API calls required for realtime instead of batch operation, ported our classifier code to run in this structure, set up a moderator user on Reddit for the bot (the choice of username was provided by the community), and a test subreddit, implemented moderation API calls, and tested it extensively before deploying it to the r/Malaysia subreddit.

Within the first two hours of being in live operation, the bot correctly detected and removed a self-promotional post, and has made not yet made an incorrect call as of the time of this contest submission.

# Interactive Demo

I have configured the r/HelloPulisBotTesting subreddit for any reader who would like to try out the bot in action. If you submit a video to this subreddit, within five minutes you should receive a private message informing you of the classification decision, and letting you know the weights it calculated for each feature of your post and video.

Some things you can try:

- Examine the removals made by the bot by viewing its comment history.
- Using an established Reddit account, post a video to the subreddit which has been on the top of r/Videos and has a high view and subscriber count.
- Create a throwaway Reddit account and post a recently-uploaded video to the subreddit (please ensure your channel's subscriber count is not hidden!)

I have also configured Reddit user with locked-down moderator permissions for the test subreddit, which readers can use to inspect the internal workings of the bot on test submissions:

- Log in as u/HelloPulisBotAuditor with password "CS109rocks"
- Visit the moderation log of r/HelloPulisBotTesting to see when the bot fires.
- View the bot's internal data store to examine its weights and feature bucket configuration.

# Discussion and Future Work

This project has produced several initial results which may be a starting point for future work.

## Toxicity scoring

It should be possible to refine toxicity scoring such that it may be calculated quickly live and used in the input to more advanced moderation classification tasks, such as directly acting on user reports related to hate speech and offensive content.

On first glance, I found it surprising that toxicity may be estimated from user report counts alone: If this is the case, why is human moderation required? But on further reflection, the reason this is effective at all is that we have done our job in creating a strong framework of norms and rules, communicated this to the community over time, and secured buy-in. This role is necessary irrespective of the level of automation we are able to apply to day-to-day policing.

## Sleep disorders

We have shown that toxic users in our community tend to have sleep patterns that are relatively heterogenous to the general population. This appears to be a novel research finding: while there is research into personality traits of online trolls and the relationship of mood to trolling (terms used in research in a similar way to our definition of toxicity), I was not able to find any research taking an epidemiological view of problematic online behavior. We may be in a good position to examine this area of research further.

My sense is that our current result may actually significantly underestimate the reality of sleep disturbances in the toxic user population. Our approach bundles in users that are toxic on the broader site, but do not participate in our community enough to have been tagged as toxic, which means that their sleep data is misclassified. We examined the dataset by hand and quickly found several extremely toxic when sorting by poor sleep quality. Sharper tools may also exist for estimating sleep quality, such as calculating variance statistics of nighttime online activity, and linking results to medical diagnostic criteria.

An especially interesting next question comes to mind as an application of this finding. If we can detect poor sleep online at scale, can we also perform interventions which help users with their sleep in a similar way? One could design a program which reached out to users at scale, informed them of medical risks, and made it easy to access resources and progress tracking. This unusual access to an unhealthy population could provide improvements to public health at minimal cost.

# Probabilistic Classifiers

This tool appears to be broadly applicable to the field of moderation at scale. Well-managed online communities design their rules to be clear, easy to interpret and easy to enforce – and produce large feeds of classification data in doing so. This seems like an ideal scenario for classification algorithms; I now see multiple areas where this tool might be applied within my subreddit alone.

The classifier we have developed is expected to save only 17 hours per year of moderator effort, but time savings scale linearly with community size. I believe a solution at scale should exist such that moderator energies might be directed towards higher-impact community development efforts.

We learned that Naïve Bayesian classifiers are surprisingly effective, even when compared against cutting-edge algorithms.

Further work on the self-promotion classifier might include adding more features to improve accuracy (such as providing user annotations data), and configuring it to improve its weights over time by watching the moderator activity feed.

# Conclusion

This project confirmed that there appears to be significant low-hanging fruit in applying modern Computer Science tools to the field of online community moderation.