MODERATION IN DIFFERENT COMMUNITIES ON REDDIT – A QUALITATIVE ANALYSIS STUDY

Iris Birman

A Thesis
Presented to
The Academic Faculty

In Partial Fulfillment
of the Requirements for the
Research Option in the
College of Computing

Georgia Institute of Technology

Abstract

The purpose of this study is to analyze moderation of different subreddits with varying degrees of moderation on Reddit. Understanding differences between how and to what extent rules, which are specified in the community norms section of each subreddit, are enforced showcases the characteristics of the subreddit itself as well as the degree to which freedom of speech is allowed. How strictly rules are enforced, in addition to the objectivity and harshness of the rules themselves, is a topic of comparison. The way moderators interact with each other, the tools and enhancements they use to aid moderation, such as Automoderator, and the way they make difficult decisions are also aspects of the analysis. This research will aid not only in understanding the moderation of Reddit subcommunities, or "subreddits," but will also help explain the workings of online communities in general. The tangential fields of sociology, industrial psychology, politics, and science will be interested in the results. The methodology of the paper involved ascertaining which of the top 100 subreddits by subscriber count are the most and least heavily moderated. Then, qualitative interviews were conducted with 15 Reddit moderators as well as the creator of the Automoderator tool, Chad Birch. The findings from these interviews are elaborated on in this paper in relation to how they answer our inquiries.

Introduction

The proposed research focuses on various sub-communities, or "subreddits," within the large online community of Reddit. Reddit is a particularly interesting site because of the variety of subreddits that exist on almost any topic imaginable. Identifying community norms, identity, and characteristics of contrasting subreddit in terms of how they are moderated is one of the goals of this project. The study will use data retrieved about moderation and deleted posts in various subreddits to better understand how effective moderation is. The objective is also to learn

about the practices and experiences of participants who are moderators in such communities. It is crucial to understand how different kinds of moderation, styles, and designs influence moderation, and how moderation can showcase the subreddit's identity accurately. Ascertaining the differences between what kinds of posts are tolerated on different subreddits helps understand the communities' identities as well as the types of users who typically flock to these subreddits. It is also important to identify and analyze the difficulties of moderation work. The information and conclusions of this research can then be used to design more effective tools for Reddit in addition to improvements to Automod, by taking into consideration the perspective and experience of the moderators.

This allows one to understand how effective moderation is in general, and how there are indeed many harassment issues that vary in severity from subreddit to subreddit, depending on their specific community norms. For example, one would expect moderation to be stricter in a more sensitive community, such as /r/mentalhealth, than a subreddit that allows for NSFW (not safe for work) posts. The research study is expected to bring more data and information to this matter, as currently there is some conflicting information regarding how safe different subreddits are. On the one hand, there is research showing that commenting by anonymous users on subreddits for users struggling with the same health problems has been therapeutic (De Choudhury and De 2014). On the other hand, there is a plethora of past and current research on online harassment and bullying in general on the Internet (Lapidot-Lefler and Barak 2011; Golder and Macy 2014).

Previous research studies have come to several conclusions, which include conflicting results (De Choudhury and De 2014; Lapidot-Lefler and Barak 2011). First, it is interesting to note a "crowd" mentality amongst users, as they are more likely to upvote or downvote a

comment if it has been done so before by another anonymous user (Gilbert 2013). Another finding of this research is that more than half of the most popular posts containing links on Reddit have been posted earlier but did not receive recognition when they were initially posted (Gilbert 2013). Social feedback is very important on Reddit because the number of upvotes a post receives helps distinguish it from other posts, thereby making the post popular on the site.

It has also been found that in weight loss communities on Reddit, users are more likely to return to the community when receiving positive feedback on their first post (Cunha et al. 2016). This finding is consistent with social psychology research that finds an improvement in health and wellness to correlate with others' positive feedback or support of the overweight person (Wang, Pbert, and Lemon 2014). If someone struggling with mental health problems expounds upon their issues in any of Reddit's pertinent communities, they are likely to receive emotional support from people suffering through the same issue, whose advice has been found to be "more involving and emotionally engaging" than typical online or face-to-face interactions for those who are struggling (De Choudhury and De 2014).

In contrast with the preceding research studies, anonymity online has also led to much anger and bullying being moved to these sites, especially specific subreddits known for hate speech. The online disinhibition effect is defined as the abandonment or reduction in normal social norms found in regular face-to-face conversations when going online or using any forms of technology that anonymize the user's identity. In analyzing which factor contributes most to this effect – among the three factors of anonymity, invisibility, and lack of eye-contact, it has been concluded the lack of eye-contact is the largest contributor to the online disinhibition effect (Lapidot-Lefler and Barak 2011).

Online platforms such as Reddit have been criticized for their moderation practices. Recently there have been controversies like those related to fake news, ISIS beheadings, and war brutalities (Allcott and Gentzkow 2017; Kelion 2013; Scott and Isaac 2016). They raise questions over how platforms sometimes fail to sufficiently remove disturbing material while censoring material that is important to discussions in the public sphere.

Although it is fairly easy to criticize moderators for making mistakes in their decisions, it is important to note that making the correct moderation decisions can be difficult. While everyone agrees that content such as hate speech and threats should be removed, it is difficult to judge what the appropriate course of action should be in many instances (Jhaver, Chan, and Bruckman 2018). For example, Facebook often takes down images of war-torn areas posted on the site, following its policy of removing violent and disturbing content. However, many users protest the removal of such images because they believe in being transparent about what goes on in warfare. Also, as social media websites grow in popularity, they face the challenge of managing content at scale, and it can be prohibitively expensive to make nuanced moderation decisions about millions of posts every day.

Importantly, we do not know much about the mechanisms driving many moderation decisions. The process of moderation can be opaque and hard for outsiders to understand. With respect to Reddit, it is currently-unclear how moderation decisions are made on various subreddits. This research study gathers and analyzes data on specific subreddits because there is not enough current research on the matter. More precisely, this research will fill the gap by generating more concrete data and evidence on moderation rates and data differences between what kind of language is acceptable in different subreddits, and what causes moderators to make certain decisions in uncertain environments. The work will also allow one to comprehend the

hierarchical structure of moderation in these communities, which tends to vary from subreddit to subreddit. This involves understanding how moderators interact with each other, and how they incorporate Automod into their moderation work.

Most subreddits rely on an Automod, which depends on a script they write that has moderating privileges just like any regular human moderator has. There is a wide range of actions the Automod may take: it can delete posts or comments if it sees a specific phrase or word that the human moderators had encoded would mean deletion, for example. It can also be set to automatically approve comments from moderators or very active users that the moderators trust no matter how many reports, sometimes jokingly, are sent in. Automod is useful because it alleviates the burden on the human moderators. It can do some of the tedious work like removing comments containing curse words or certain phrases that are banned in the subreddit.

Seeing which hierarchical structures of moderation, involving tools and helpers like Automod, are most effective in maintaining the community norms is useful in industrial and organizational psychology theory as well. This research may also serve as constructive advice for other subreddits and even other websites besides Reddit that heavily moderate their discussion forums. Currently, there is a lack of a high-level understanding of experiences and decision-making processes. This research seeks to solve the issue, providing informative insights and ideas for future exploratory work.

Literature Review

To better understand the true effectiveness of moderation on Reddit's subcommunities, or "subreddits," it is important to conduct this proposed research study by analyzing the deleted comments themselves, thread structures, and statistical evidence regarding deletions in these various online communities. The results of the study are going to be obtained through both statistical and textual analysis of social behaviors in subreddits as well as informative qualitative

interviews. This study is necessary because of extant research providing conflicting results regarding the true usefulness of subreddits. For example, looking at current research is not enough to understand if these communities serve to benefit or worsen the social interactions between users on this site as well as the overall psychological states and behavioral manifestations of the users (De Choudhury and De 2014; Lapidot-Lefler and Barak 2011).

Gilbert (2013) discusses the idea of "underprovision" on Reddit – that is, when too many people rely on others to make decisions or contribute to upvoting or downvoting on Reddit rather than doing the voting firsthand by themselves. This study finds that underprovision is widespread on Reddit and discusses some possible reasons behind it. Also, the study shows that Reddit apparently overlooked 51.52% of the most popular links on Reddit the first time they were submitted to the site: these popular links had been posted earlier without getting any recognition back when they were posted in the beginning (Gilbert 2013). The methods and statistical analysis and evidence were done thoroughly on 9,370 Reddit links, and this study is applicable to Reddit users in understanding crowd mentality online and indeed how similar it is to crowd mentality in person.

Toxicity Online

Lapidot-Lefler and Barak (2011) analyze the toxicity of online comments. They look at manifestations, as assessed by judges through textual analysis and self-reports, of what's known as the toxic online disinhibition effect. Toxic online disinhibition effect is defined as reducing or abandoning normal social norms or customs usually found in face-to-face interpersonal conversations when going online or using forms of technology that include anonymity of the user. This research article particularly questions which of the following three factors contributes most and to what extent they all contribute to this effect: anonymity, invisibility, and lack of eye-

contact. The study concludes that lack of eye-contact is the biggest contributor to the negative effects of online disinhibition effect (Lapidot-Lefler and Barak 2011).

The experimental design included a rather small number of participants: 142. This study which concludes that lack of eye-contact is the main reason for the disinhibition effect that we see is distinct because it implies that if there were eye contact present – if this were to be a face-to-face conversation between one of the members and a person whom they would be typically insulting online – that this inappropriate behavior would not occur in this face-to-face conversation. The results are presented in a coherent manner and organized well and succinctly such that the reader can understand them.

Supportive Posts and Comments

Reddit weight loss subreddits offer another distinct vantage point from which to view the efficacy of Reddit's subcommunities. In a 2016 paper by Cunha et al., the researchers asked if social feedback received on a user's very first post in the community affects the chance that this user will continue being a part of the community, whether social feedback on a user's later posts is as impactful as feedback on their first posts, and what the characteristics of initial posts are that attract more social feedback than average. They find that users are more likely to return to the community when receiving positive feedback on the first post and that feedback on later posts does not have as much of an effect as feedback on earlier posts on the user's timeline in this community (Cunha et al. 2016).

The research results achieved came from a data collection period spanning five years. The researchers obtained the data using PRAW (Python Reddit API Wrapper), which gives the researcher access to Reddit's official API. They ran their research on a total of 70,949 posts and 922,245 comments, which were linked to a total of 107,886 unique users. Most of these users

had not actually posted themselves, but were instead active commenters: only 36.1% wrote a post, but 93.6% wrote at least one comment. This is in line with typical Reddit behavior, where there are often more commentators than active posters. In this article, statistical analysis is combined with qualitative analysis in a coherent manner. The study deals with social feedback in weight loss communities online, and how that affects the original poster's longevity on the site.

A study done in 2014 by De Choudhury and De looks at how users post about their mental health issues on Reddit and how they receive considerable social feedback and support. The researchers also develop statistical and language models to quantify the types of mental health social support. Studying the results has led the researchers to conclude that anonymity on Reddit sparks rather open conversations surrounding mental health problems and that the feedback is "more involving and emotionally engaging" (De Choudhury and De 2014).

In the methods, the researchers do a textual analysis of Reddit posts and comments to find indications of depression. The methods section is very extensive and has numerous statistical evidence, charts, and tables to support the results. It is a thorough body of work, from which it is evident that Reddit can fill voids in emotional support to people who are struggling because of the support system that is easily found if a user simply navigates over to a specific community for his or her issue. The quality and depth of comments is very important.

In the proposed future research, in comparison to Lapidot-Lefler and Barak's study, it would be better to have at least more than a few hundred participants, since using more participants means the research claims will be stronger. In my research, when I collect and analyze data from subreddits that are typically known for members ranting or insulting other people, the question of anonymity perhaps leading or augmenting such behavior is very relevant.

This research also goes beyond current research by analyzing more instances of subreddits with contrasting deletion rates. The question of how moderators help create a specific community structure or identity is an important one. How are the moderators able to maintain a place on the Internet where all posts and comments are on topic, void of any spam, inappropriate, or off-topic commentary?

For this reason, it was necessary to delve first into how moderation happens on this site, and to find differences between posts that were kept and removed. Thread structures under posts, including ones posted in banned and popular communities, and the user timelines of people who get banned or whose posts are deleted, were studied. Also, studying the interactions between moderators and users of their sub was important. In such a way, this research fills the gap in understanding indeed how effective the online community of Reddit is through comprehension of different subreddit identities and norms by looking specifically at how these communities are moderated. Comparing moderation techniques between subreddits of very different moderation styles allows one to understand the functioning of different communities, to what extent they facilitate free speech, and how to potentially improve the system in the future.

Methods and Materials

This study centers more around qualitative research, though it does involve preliminary quantitative analysis.

Data Collection

The first half of this research study included quantitative data collection. Data was collected for the top 100 subreddits, as determined by the subscriber count at http://redditmetrics.com/top in July 2017. SQL queries were written and processed in Google BigQuery, which contains all past Reddit data, such as comments and posts. These queries got

the total number of comments submitted to each subreddit in the time period from June 1, 2016 to December 31, 2016. Another set of queries were run to get the total number of removed comments, and then this number was divided by the total number of comments per subreddit to get the percent comments removed per subreddit. Removed comments are defined as comments removed by a moderator, administrator, bot, or anyone who is not the user who wrote the comment itself. This data was viewed in Microsoft Excel and sorted to see which subreddits had the highest and lowest removal rates.

Qualitative Interviews

The second half of this research involved qualitative research, conducted through interviews with moderators of the most and least heavily moderated subreddits, as determined by the removal percentages. In addition, Chad Birch, the creator of Automod who currently moderates r/automoderator and used to moderate r/games, was interviewed. As described in the Introduction, Automod is a tool that helps automate moderation on Reddit using a word-filtering based approach.

The goal was to understand more about the moderators' moderation techniques and the differences in community norms from subreddit to subreddit. Three moderators per subreddit were interviewed. Interviews were semi-structured, based off an interview questionnaire guideline that can be viewed in Appendix A. The interviews lasted a maximum of 90 minutes each, during which as many of the questions on the questionnaire as possible were asked. Follow-up questions were also included at moments where it was necessary to clarify or expand upon something important or interesting the interviewee had said. Some of the interviewees answered all the questions fully in the time allotted, while others took longer and answered only

part of the questions. Interviews were conducted via Reddit's new chat feature, phone, or videoconference (Skype).

Purposeful Sampling

The subreddits chosen for the subsequent moderator interviews were picked with purpose. Three of the subreddits were in the top subreddits by percentage of deleted comments. These were r/photoshopbattles, r/space, and r/explainlikeimfive. The fourth subreddit was chosen from the 10 subreddits with the lowest deletion rates, specifically for having one of the highest mod to comment ratios. Out of the 10 bottom subs, r/announcements had the highest mod to comment ratio, but since it was somewhat unusual, the second highest mod to comment ratio, which was that of r/oddlysatisfying, was chosen instead. This meant that while the subreddit has enough labor to delete comments, its rules probably state that it should not delete much. For this reason, mods from this subreddit – r/oddlysatisfying – were interviewed. Moderators from r/politics were also interviewed, due to r/politics falling in the medium-range deletion level. To sum up, moderators from r/photoshopbattles, r/space, r/explainlikeimfive, r/oddlysatisfying, and r/politics participated in this research study. This provides a better understanding of how different types of communities choose to keep their subreddit running, and what kinds of behaviors they permit versus reprimand.

A table of participants is provided below. OS_01 shared his real name, Mitchell Creed, because he is the only remaining co-founder of r/oddlysatisfying, and takes a lot of credit for the sub's continuing success.

SUBREDDIT	PARTICIPANT	AGE	OCCUPATION	COUNTRY
r/photoshopbattles	PB_01	33	Engineer	USA
r/photoshopbattles	PB_02	Late 20's	Engineer	Australia

r/photoshopbattles	PB_03	Not available	Not available	UK
r/space	Space_01	25	Graduate student	USA
r/space	Space_02	20-25	Student	India
r/space	Space_03	33	Supply chain manager	USA
r/oddlysatisfying	OS_01 (Mitchell Creed*)	32	Student	USA
r/oddlysatisfying	OS_02	21	Student	USA
r/oddlysatisfying	OS_03	26	Audiologist	UK
r/explainlikeimfive	ELI5_01	30	Software developer	USA
r/explainlikeimfive	ELI5_02	27	Fraud analyst	USA
r/explainlikeimfive	ELI5_03	28	Data analyst	UK
r/politics	Pol_01	32	Not available	USA
r/politics	Pol_02	25	Student	USA
r/politics	Pol_03	25-29	Not available	USA
r/Automoderator	Chad Birch	34	Unemployed	Canada

* Note: This moderator chose to have his full real name included.

Analysis

Data from the interviews was transcribed and read over multiple times. Interpretive qualitative analysis was applied to all interview transcripts and field notes (Merriam 2002). The process involved categorizing the data, identifying patterns and then, grouping them into themes. The analysis first involved what is known as "open coding," where descriptive short phrases were manually assigned as codes to the data (Charmaz 2006). Examples of these codes include the following: "reversing other moderators' actions," "regretting getting too invested in arguments," and "believing that choice of keywords for Automod is important." The first round of coding was completed on a line-by-line basis so that codes stayed close to the raw interview data. 481 first-level codes were gathered at this stage.

The next step of the analysis involved more rounds of memo-writing and coding, where initial codes were compared along with the data to which they were bound. Seven key themes, such as “Difficulties of moderation work,” were generated by this process. Some of these themes will be reported in the paper; themes such as “Recruiting new moderators” and “Becoming a moderator,” were excluded in further analysis because they were generally basic and similar to each other no matter what subreddit was being discussed. Finally, connections between themes were established, contributing to how the phenomena are described in the Results, presented next.

Results

One of the main outcomes of this research was to gain an understanding of the difficulties of moderation work. Another was to understand the motivations behind moderating, despite the fact that it is basically unpaid volunteer work that can take up a lot of time and subjects the moderator to verbal harassment.

Deleted Comment Rates

The results from the preliminary quantitative analysis of deleted comment data can be found below in Table 2. Note that in the analysis, only the comments that had '%[removed]%' were those that were taken down by moderators. Comments with the '%[deleted]%' label were taken down by the users themselves, so they do not matter in this research analysis of the moderators' work. '%[deleted]%' comments are shown here, though, for informative purposes. Note that 'reddit.com' subreddit, which can be accessed at reddit.com/r/reddit.com has no new posts and comments posted within the last 6 years because the subreddit is archived and no longer receiving new submissions. But because it was a default subreddit, and people were automatically subscribed to it, it remains in the top 100 subreddits by subscriber count.

Subreddit	Number of total comments (June 1 - Dec 31, 2016)	# comments '%[deleted]%'	# comments '%[removed]%'	Fraction of comments taken down using 'removed'	As a percent
Overwatch	4,740,789	158,624	13,217	0.002788	0.2788
announcements	123,954	9,449	348	0.002807	0.2807
programming	460,667	23,852	1,930	0.004190	0.4190
todayilearned	3,137,881	175,349	19,291	0.006148	0.6148
oddlysatisfying	180,978	10,050	1,116	0.006166	0.6166
mildlyinteresting	1,162,053	69,591	7,502	0.006456	0.6456
movies	2971827	141,436	21,649	0.007285	0.7285
Malefashionadvice	585,260	23,647	5,098	0.008711	0.8711
Fitness	1,030,433	46,771	9,688	0.009402	0.9402
pokemongo	3,678,123	138,351	38,090	0.01036	1.036
WTF	1,581,287	88,853	16,965	0.01073	1.073
trees	1,116,381	53,281	12,186	0.01092	1.092
YouShouldKnow	58,442	3,207	671	0.01148	1.148
nba	4,144,860	217,812	48,640	0.01174	1.174
Android	1,291,561	50,205	15,221	0.01178	1.178
soccer	3,870,798	220,930	47,083	0.01216	1.216
comics	401,238	12,927	5,091	0.01269	1.269
StarWars	898,014	25,878	11,633	0.01295	1.295
blog	54,567	2,084	735	0.01347	1.347
nfl	4,919,847	201,608	66,299	0.01348	1.348
pokemon	6,991,014	215,687	94,379	0.01350	1.350
AskReddit	33,794,947	1,782,194	471,545	0.01395	1.395
videos	3,802,383	238,110	54,542	0.01434	1.434
facepalm	213,975	12,555	3,240	0.01514	1.514
pcmasterrace	3,530,010	148,614	53,495	0.01515	1.515
AdviceAnimals	1,759,126	90,712	27,753	0.01578	1.578
pics	4,518,373	280,734	72,695	0.01609	1.609
Unexpected	159,461	9,571	2,586	0.01622	1.622
woahdude	236,244	13,421	3,893	0.01648	1.648
atheism	812,595	32,120	13,562	0.01669	1.669
Documentaries	330,952	19,757	5,916	0.01788	1.788
Music	944,468	40,048	22,089	0.02339	2.339
gifs	2,848,806	168,968	67,052	0.02354	2.354
leagueoflegends	5,026,706	184,892	120,578	0.02399	2.399

FoodPorn	71,089	3,306	1,718	0.02417	2.417
gameofthrones	1,031,543	39,473	25,294	0.02452	2.452
books	902,488	29,662	22,899	0.02537	2.537
lifehacks	55,807	2,771	1,518	0.02720	2.720
OldSchoolCool	577,366	37,232	16,012	0.02773	2.773
politics	14,391,594	680,783	402,509	0.02797	2.797
interestingasfuck	412,479	22,328	12,096	0.02933	2.933
europe	1,159,180	65,932	36,728	0.03168	3.168
BlackPeopleTwitter	810,845	47,675	26,910	0.03319	3.319
aww	1,293,794	65,523	43,170	0.03337	3.337
gonewild	1,623,942	114,362	54,854	0.03378	3.378
Showerthoughts	2,192,293	116,962	78,896	0.03599	3.599
LifeProTips	722,650	36,448	26,700	0.03695	3.695
television	1,063,465	52,751	39,515	0.03716	3.716
funny	3,995,506	234,090	150,640	0.03770	3.770
wholesomememes	56,410	2,203	2,179	0.03863	3.863
reactiongifs	243,966	14,349	9,568	0.03922	3.922
tattoos	129,547	5,650	5,087	0.03927	3.927
technology	1,109,229	49,394	44,723	0.04032	4.032
DIY	426,670	17,325	17,336	0.04063	4.063
worldnews	4,871,009	288,983	198,401	0.04073	4.073
EarthPorn	248,664	10,615	10,312	0.04147	4.147
nottheonion	721,164	37,938	29,929	0.04150	4.150
gaming	4,073,084	162,354	174,614	0.04287	4.287
Frugal	230,431	10,799	10,072	0.04371	4.371
Jokes	814,620	46,286	35,714	0.04384	4.384
4chan	440,282	19,497	19,831	0.04504	4.504
GetMotivated	192,283	11,314	9,190	0.04779	4.779
cringepics	195,137	11,669	9,600	0.04920	4.920
InternetIsBeautiful	67,978	3,625	3,426	0.05040	5.040
food	641,472	29,494	33,488	0.05220	5.220
sports	683,287	33,625	38,184	0.05588	5.588
IAmA	1,072,040	55,686	60,712	0.05663	5.663
nsfw	225,183	13,098	14,193	0.06303	6.303
relationships	2,334,077	148,564	149,237	0.06394	6.394
sex	756,807	39,926	48,424	0.06398	6.398
news	10,194,316	565,330	654,906	0.06424	6.424
OutOfTheLoop	231,192	10,837	16,787	0.07261	7.261
ImGoingToHellForThis	296,003	18,257	21,550	0.07280	7.280

tifu	1,186,683	56,044	89,714	0.07560	7.560
creepy	421,617	18,193	31,938	0.07575	7.575
Art	369,522	19,264	28,393	0.076837103	7.684
WritingPrompts	380,023	9,592	29,590	0.077863708	7.786
personalfinance	1,074,779	48,378	84,453	0.07858	7.858
me_irl	587,874	26,263	46,308	0.07877	7.877
philosophy	221,149	10,127	20,531	0.09284	9.284
RealGirls	106,650	9,040	10,505	0.09850	9.850
Futurology	728,561	28,686	76,970	0.1056	10.56
explainlikeimfive	964,821	65,739	105,846	0.1097	10.97
Games	1,371,882	48,024	151,758	0.1106	11.06
TwoXChromosomes	774,268	46,615	87,522	0.1130	11.30
listentothis	146,319	3,988	17,616	0.1204	12.04
gadgets	256,009	10,206	31,415	0.1227	12.27
UpliftingNews	287,880	16,480	35,876	0.1246	12.46
nosleep	292,002	8,919	39,014	0.1336	13.36
dataisbeautiful	474,685	22,203	64,689	0.1363	13.63
bestof	312,191	13,188	46,738	0.1497	14.97
history	320,995	10,347	53,418	0.1664	16.64
NSFW_GIF	123,777	8,277	21,580	0.1743	17.43
space	795,186	24,499	146,832	0.1847	18.47
HistoryPorn	111,465	4,725	25,711	0.2307	23.07
askscience	438,243	12,610	136,242	0.3109	31.09
science	1,084,955	36,041	372,340	0.3432	34.32
AskHistorians	120,715	3,824	41,956	0.3476	34.76
photoshopbattles	300,369	26,434	112,906	0.3757	37.59
reddit.com	0	0	0	0	0

Challenges of Moderation

A common theme or topic of concern among moderators is how to deal with a post that hits r/all even if it violates community norms. On r/oddlysatisfying, a post like this will still get removed even after appearing on the r/all page “because it either got mistakenly approved by a mod or no mods were active at that time. We still have to remove those posts, which upsets some users, but it is only fair to everybody if we do so rather than making exceptions,” said OS_01.

Also, it is problematic when a post hits r/all and causes an influx of new commentators who do not know the rules or norms of this particular sub. “A lot of times what happens . . . the post will hit r/all and then people will pour in from outside the sub, who aren’t familiar with the rules or never posted or commented before,” said participant Space_03.

Some moderators also experience difficulties searching for old posts. A moderator on r/oddlysatisfying said that, "With reposts, I could be very sure that something was posted in the last 2 months (one of our rules is to remove things posted within the last 2 months), however I may not find it in the search engine. Now, that is either because of reddit's poor search engine, or it had different keywords, or it was posted to another subreddit" -- *OS_03*. Another moderator from the same subreddit said that it is not hard to make decisions on what to remove or keep: "Our rules are pretty simple to enforce . . . I think the hardest decision is about our 'clickbait title' rule. If a user doesn't describe what is happening in the post with their title – such as 'check this out' or 'so satisfying...' -- I will remove it and the user generally resubmits with a descriptive title. Because we maintain the rule that nearly anything can be satisfying, we remove posts very rarely if they fit all other rules" -- *OS_02.*

With respect to r/oddlysatisfying, OS_02 notes that the subreddit specifically has “a rule against defining what exactly oddlysatisfying is,” so the poster’s subjective view of what is oddlysatisfying is not disputed by the moderators unless is violates an actual community rule like the descriptive title rule mentioned earlier. The issue is that other users of the sub will often report posts that they think are not satisfying enough. However, the moderators cannot change remove it because of anyone’s subjective opinions, so nothing can really be done about this subjective opinion situation.

Another moderator complained about Reddit's interface: "The moderation interface requires more clicks than is strictly necessary, the mod mail (until recently) was cluttered and difficult to use, there's still a very little to no ability to track users between infractions, and there exists no ability to determine if someone is 'ban evading' by creating a new account after being banned from an old one unless they volunteer that information. Those are the things that bother me, specifically, but I'm sure there are a great deal more that I've become accustomed to over time" -- *ELI5_02*.

Sometimes, if a user's comment or post is removed, they will message the moderators with accusations or verbal harassment, some of it personally aimed at the moderator who removed the user's post or comment. "When we remove lots of off-topic or trolling comments, we're often accused of 'censorship,' or when aggressive political comments are removed we're accused of being in the pocket of the opposite side of whatever group they were supporting. It's amusing to be a 'libtard' and a 'Trump shill' in the same day" -- *Space_01*.

Reasons for Continuing Moderation Work

One may ask themselves the question, “Why would you continue to do the job of a moderator if you receive near constant harassment and almost no praise?” Surprisingly, it is still an awarding experience. One moderator specified that most of the angry messages he receives are actually from Reddit users that are not active on his sub; meanwhile, he receives an overwhelming amount of praise from the more active participants in the community.

> “We found when we track the people who complain about moderation, versus the people who compliment the moderation, that people who are more involved in the sub are more likely to want more active moderation. Like recently there was a big event where a bunch of people posted publicly about the moderation in the sub, and I tracked who posted

> negatively about it and who posted positively about it. I went back and looked over all the user comment histories, and of the people that had posted negatively about moderation, half of them had never before posted in space. And then the average, you know, the average number of comments from each user who posted negatively in that sub was like 1.5 or something like that, something low. And then the average number of comments by people who had posted positively was like 25. And a lot of them, it's clear that they had been commenting more but you know, after a certain point, Reddit doesn't store the comments anymore so you can't see them." – *Space_03*.

This idea was reinforced by another mod of the same sub, r/space, "The users that complain in those cases have often never posted in the subreddit before. And they came in when a popular post reached r/all" -- *Space_01*. With respect to r/oddlysatisfying, the only remaining co-founder who is still a moderator, OS_01, said, "We have gained a great reputation within the community as having a friendly atmosphere overall for both users and mods."

Space_03 also clarified that his passion for moderation does not stem from a desire to control things: "It's not about control. It's not like I want to exercise my little reign of power. It's just going into the sub; there's a bunch of stuff just kind of like off-topic and it's not really salient, and I'll deal with it for 5 minutes and then I'll leave, and it looks nicer than when I found it."

Different Moderators' Approaches

Moderators can also have rather different approaches to moderation: one moderator explained, "There are some mods that are aggressive to users and will not reverse a decision on a post if questioned or asked about it while another mod might take a second look and agree with the user and re-approve the post. There are a lot of mods that don't even go into the unmoderated

queue, and just look at user/automod reports as well" – *OS_02*. However, such unenthusiastic behavior is atypical when regarding the majority of moderators. These individuals volunteer their time and effort mainly because they just really enjoy what they do.

Another moderator stated that he is usually lenient toward borderline posts – or posts where it is not clear whether they should be removed. "Sometimes I try to engage people that complain about the rules, or let certain political or unserious comments near the margin of what's appropriate slide. I sometimes find that another mod has removed a comment I was hesitant about, and that's usually fine with me" – *Space_01*. Most moderators such as this one expressed similar ideas about how they usually do not reverse the decisions of other moderators. Every moderator typically has the same privileges, and their judgment call is usually respected by the other moderators.

How Automoderator Works

The first version of the Reddit Automod was created in January 2012 by Chad Birch while he was a moderator on r/gaming. He noticed that many of his tasks as moderator were very repetitive and easy, such as searching for keywords in each post and comment that meant it should be removed. Chad built the very first Automod as a bot set with conditional actions, applying these conditions to any newly posted content and then performing the configured actions if those conditions were met. Regular expressions, which allow for defining patterns and words, were used in the condition statements. Regular expressions themselves are defined as special text strings for describing specific search patterns, which are used when searching large bodies of text to find words or phrases that match the patterns (Thompson 1968).

For example, one subreddit had configured Automod using the following regular expression to catch and remove many homophobic slurs, in efforts to deal with hate speech:

(ph|f)agg?s?([e0aio]ts?|oted|otry)

This single expression catches many slur words such as “phagot,” “faggots,” and “faggotry.” The regular expression includes conditions that can be combined or inverted in any manner. Looking at this example, we can see that the first part of the regular expression tests for whether the word starts with “ph” or “f,” which is indicated by the OR operator |. There are different kinds of operators and symbols that are used to construct these expressions. Using the conditions defined by regular expressions, Automod develops various capabilities such as accepting posts and comments from known, trusted users, removing comments containing banned phrases, or auto-approving submissions from users whose account age and karma points are higher than some threshold values decided upon by the moderators. As participant Space_02 put it, “for the parts that can be automated, the bot does most of the job. Like we can tell it to autoremove links from a certain domain if we know a website a crap or ask the bot to remove all one-word comments since they don’t contribute to the discussion meaningfully.”

The Role of Automoderator

Moderators use the automated mechanism of Automoderator, or Automod for short, to help ease their workload. “I think about 2 in 10 posts are automod removing a post because of the new user spam protection . . . We also have protection against imgur album submissions because spammers will embed a phishing URL in the description. Basically, if a user’s account is under X days old or has less than X karma it will be automatically placed in the report queue to make sure it’s not a spammer” – *OS_02.* Here, notice that this subreddit’s first round of defense against spammers involves Automod removing or reporting issues that may or may not be reversed by the human moderators.

But Automod was not always a big part of the moderation work. In the beginning, many moderators had resisted using Automod: “There [were] also just a lot of moderators who were

just worried about having a bot moderate their subreddits. Since it had always been a manual process until that point, it was just a really big switch in the way they thought about moderation – to automate a lot of it" – *Chad.* At first, the Automod was also rather inconvenient to run. Chad set it up for a few select moderators, who had to either set up and run their own instance of a rather complex bot programmed in Python or contact Chad every time they wanted to make changes to the configurations. So Chad continued to improve Automod, culminating in the release of a new version of Automod in May 2013 that was self-configurable by moderators through Reddit's wiki system (Birch 2015). The newly-implemented ability to directly influence the rules of Automod quickly made this tool popular among moderators, so more and more subreddits started using it. The increasing implementation of Automod among various subreddits was followed by Reddit's official adoption of the Automod tool in March 2015.

Each subreddit now has its own wiki page, accessible only to its moderators, for configuring the rules of their respective Automod. Any changes to the wiki immediately go into effect. The wiki page also keeps track of all edits, so it is easy to see which moderators make which modifications to the rules of Automod. This system establishes accountability for one's moderation actions and allows for reversal of any changes that are made.

Challenges of Using Automoderator

Automod is lauded by moderators for the improvements it has on efficiency and for dealing with a fair amount of boring, repetitive work such as searching for banned words in posts and comments. But using Automod has its own challenges. First, Automod requires those who code it to understand regular expressions, which means that some moderators must learn new material that may be unrelated to their actual profession or interests. One participant, OS_01,

said he has “no clue how to work it, but one of our other mods is pretty savvy in regard to programming it.”

Pol_01 said, “it’s not necessarily user-friendly . . . it almost entirely functions on regex, and its own little quirks and syntax to implement things, so it can take some time for people to get decent at using it.” In efforts to resolve and prevent misunderstandings that arise from not knowing regular expressions extensively, it is important to have moderators on the team that know how to write and make changes to regular expressions. Pol_01 clarifies that “it’s very useful to have people on your team that know how to use [Automod]. Other teams might want a moderator that knows CSS for help in designing the subreddit, or another common skillset teams look for is one’s ability to write bots to help aid in moderating tasks.”

When moderators are unfamiliar with regular expressions, this can also lead them to write rules that are too all-encompassing, or too broad. These rules end up removing perfectly appropriate content that was not supposed to be removed. If moderators do not check the actions of Automod by tracking what has been automatically removed, they will not know if something was removed that should not have been, unless of course they receive mod mail from the user whose post or comment was removed unintentionally. “Maybe in the range of 40-50% of the moderation actions in the sub are done by Automoderator. So a lot of times, when people come back to us on something, it will be over something that was done automatically” – *Space_03.*

There is also concern that many moderators prefer to minimize false positives rather than false negatives, meaning that they focus on ensuring that Automod catches as many inappropriate posts and comments as possible, but do not pay enough attention to whether Automod removes content that should not have been removed. The following quote exemplifies one of many such situations: “I think one of the things that bothered me before I was a moderator

and complained about a lot was the false positives. Like, half the time, the Automoderator would have automatic comment removals if your comment had the word "homo" in it. But homo is not just a gay slur; it's also the genus of human beings – *homo sapiens* – so if you would write anything about, writing a comment using the word *homo sapiens* in it, your comment would be removed." – *Space_03.*

Another challenge of Automod involves understanding the tradeoffs between reducing human moderator intervention and avoiding overly restrictive moderation brought upon by strict rules encoded into Automod. For instance, many subreddits have a rule configured on Automod that removes posts and comments from new users, which helps in dealing with trolls who create multiple accounts or keep creating new accounts after getting banned from the community. However, such rules inadvertently mistreat regular non-troll users who are just new to the site and want to share legitimate content. This situation demotivates new users from continuing to post or comment on the subreddit:

> "That sort of thing is really harmful to legitimate new users where they create an account on the site, they make comments and then, maybe they all get auto-removed, and the user doesn't even know. They just think that they registered on Reddit, but Reddit's boring because they left some comments, and nobody replied or even voted on their comments." – *Chad.*

Future Work

Due to the nature of this research as a mainly qualitative study, in the future it would be appropriate to formulate a mixed methods study based on the results of this one. Striving toward more mixed methods, thereby by definition incorporating both qualitative and quantitative methods, will provide more quantitative results on topics delineated or emphasized in this paper.

This way, any trends or interesting findings of this research can be bolstered by numerical analysis done in the future. Specifically, we would need to evaluate the statistical significance of Automod and the exact percent of comments and posts it removes in each subreddit. In the future, it would also be beneficial to the moderators if new types of Automods or new changes to it are created based on the criticism and room for improvement mentioned by the moderators interviewed in this research work. Then, they can be tested for efficacy.

References

Allcott, Hunt, and Matthew Gentzkow. 2017. “Social Media and Fake News in the 2016 Election.” *Journal of Economic Perspectives* 31, 2: 211-236. https://doi.org/10.1257/jep.31.2.211

Birch, Chad. 2015. “Moderators: AutoModerator is Now Built into Reddit – New Syntax and Functionality.” *Rebrn.com.* http://rebrn.com/re/moderators-automoderator-is-now-built-into-reddit-new-syntax-a-217992/

Charmaz, Kathy. 2006. “Coding in Grounded Theory Practice.” *Constructing Grounded Theory*: 42-70.

Cunha, Tiago, Ingmar Weber, Hamed Haddadi, and Gisele Pappa. 2016. “Effects of Social Feedback in a Reddit Weight Loss Community.” *ACM Digital Health 2016.* https://arxiv.org/pdf/1602.07936.pdf

De Choudhury, Munmun, and S. De. 2014. “Mental Health Discourse on Reddit: Self-Disclosure, Social Support, and Anonymity.” *ICWSM ’14.*

Gilbert, Eric. 2013. “Widespread Underprovision on Reddit.” *CSCW ’13.* http://comp.social.gatech.edu/papers/cscw13.reddit.gilbert.pdf

Golder, Scott, and Michael Macy. 2014. “Digital Footprints: Opportunities and Challenges for Online Social Research.” *The Annual Review of Sociology* 40: 129-152.

Jhaver, Shagun, Larry Chan, and Amy Bruckman. 2018. “The View from the Other Side: The Border Between Controversial Speech and Harassment on Kotaku in Action.” *Under Review at Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems.*

Kelion, Leo. 2013. “Facebook Lets Beheading Clips Return to Social Network.” *BBC News.*

http://www.bbc.com/news/technology-24608499

Lapidot-Lefler, Noam, and Azy Barak. 2011. “Effects of Anonymity, Invisibility, and Lack of Eye-contact on Toxic Online Disinhibition.” *Computers in Human Behavior* 28: 434-443.

McGillicuddy, Aiden, Jean-Grégoire Bernard, and Jocelyn Cranefield. 2016. “Controlling Bad Behavior in Online Communities: An Examination of Moderation Work.” *Thirty Seventh International Conference on Information Systems, Dublin 2016.*

Merriam, Sharan B. 2002. “Introduction to Qualitative Research.” *Qualitative research in practice: Examples for discussion and analysis* 1.

Pater, Jessica, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. 2016. “Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms.” *Proceedings of the 19th International Conference on Supporting Group Work (GROUP ’16),* 369-74. ACM, New York, NY, USA.

Scott, Mark, and Mike Isaac. 2016. “Facebook Restores Iconic Vietnam War Photo It Censored for Nudity – The New York Times.” *The New York Times.* https://www.nytimes.com/2016/09/10/technology/facebook-vietnam-war-photo-nudity.html

Thompson, Ken. 1968. “Programming Techniques: Regular Expression Search Algorithm.” *Communications of the ACM* 11, 6: 419-422. https://doi.org/10.1145/363347.363387

Wang, M., L. Pbert, and S. Lemon. 2014. “Influence of Family, Friend and Coworker Social Support and Social Undermining on Weight Gain Prevention among Adults.” *Obesity.*

Appendix A: Interview Guidelines

Interviews follow a semi-structured format. Questions cover these general areas:

- Tell me about your Subreddit. How did you first hear about it?
- How did you become a mod?
- Tell me about modding your sub.
    - What is a typical day like?
- How often do you find it difficult to judge how to moderate?
    - Tell me about instances when you have found it difficult?
    - How do you resolve difficulties?
- Tell me about the posting rules.
    - Can you explain each one?
    - Which rules are easy to judge, and which ones are hard?
- What kind of conversation are the mods trying to foster? How do the rules help?
- Do the mods ever ban anyone from the sub? Tell me about it.
- Do different mods have different approaches?
    - Do you ever disagree with the decisions of other mods? If yes, tell me more about it: why do you disagree?
- Do you know how the sub's rules originated and who came up with them? How have they changed over time?
- Can you tell me about the most recent rule change? What led to that rule change?
- Do you use an Automod, and how does it fit within your workload? What percent of comments and posts does the Automod take care of? What percent would be false positives, in your opinion?
- Do you think the mod team ever makes mistakes?
- Who do you think makes the best kind of mod? A person who is a very active participant but lacking mod experience, or the opposite – someone who has a lot of mod experience on other subs but is not that active on this sub?
- Have you seen users complain about moderation on your sub? What do they complain about?
    - Do you think that complaint is legitimate? Why or why not?
    - Do users' responses to moderation affect how moderation is done on your sub?
- Is there anything about the sub that isn't working the way you would hope?
- Appropriateness versus free speech can be seen as a tradeoff. How do you see the balance for your sub?
    - What do you define as appropriate on your sub?
- What tools do the mods use to discuss moderation among themselves?
- How do moderators handle disagreements about removals or bans?
- Are you a mod for any other subreddits?
    - What are the other subreddits you moderate?
    - How do you balance your workload?
    - How are your mod strategies on those subreddits different from the ones you use on this subreddit?
- Are you a regular user on any sub where you are not a mod?

    - Do you like moderation on that sub? Why or why not? How is it different from moderation on your own sub?
- Can you tell me about yourself?
    - Age?
    - Gender?
    - Occupation?
    - Where do you live?
    - Political affiliation?