Analysis of Moderation in a Reddit-like Network

Class report for
Complex Social Systems: Modeling Agents, Learning and Games

Fabian Blatter, Frederick Wolff, Max Osterried, Yuri Simantob

Eigenössische Technische Hochschule Zürich, Switzerland
{fblatter, wfrederick, mosterried, ysimantob}@student.ethz.ch

Supervisors:
Prof. Dr. Dirk Helbing
Dr. Nino Antulov-Fantulin
Dino Carpentras

Abstract

We live in a world where polarisation, fake news and regulation are terms quickly associated with social media. As the internet revolutionizes our world, social media widely changes the way we communicate. The filters applied in these platforms can give people a voice that need to be heard, but can also lead to echo chambers or be abused to spread or enforce an opinion of people in power. In this paper we analyse a Reddit-like network and the effects of a moderator that tries to minimise extremist opinions in the community by methods currently used by states and platforms. Our model is able to show that current moderation techniques work on a vast majority of the users, but can also foster extremist behaviour. As social activity shifts more and more to the online world, we deemed it important to assess the outcomes of moderating these platforms. To the knowledge of the authors this is the first attempt to analyse the effects of moderation on opinions in a social media network based on techniques applied in reality.

Contents

List of Figures

List of Tables

1 Introduction

Censorship or moderation of social media is not a new topic in our society [16]. As the internet naturally arose as a free and unregulated place, it slowly developed more and more to a platform where all kinds of opinions and groups form. Social media has become central to modern political discourse[7], and as a result, these platforms have become increasingly polarized and divisive[9]. This polarization is fueled by the ease with which people can access and share information and opinions online and the resulting phenomenon of echo chambers[8]. On social media platforms, people tend to follow and interact with peers who share their views, leading to the creation of ideologically homogeneous groups[14, 9]. This can result in the amplification of certain viewpoints, while others are drowned out or marginalized. The rise of social media has also led to the formation of extremist groups, many of which use these platforms to spread their messages and recruit new members. Such groups can range from far-right or far-left political movements to extremist religious or ideological organizations. The anonymity and reach of the internet make it an ideal platform for these groups to spread their ideologies and mobilize followers[10].

Social media and the real-life impact it has is therefore a highly controversial topic. On the one hand it enables free speech and gives people a voice that wouldn't be heard if these platforms didn't exist, which can be crucial in times where people are suppressed and have to raise their voice for their freedom. On the other hand it boosts polarisation and extremism, weakening democracies by opening the doors for populism, fake news and formation of bubbles, which prohibit consensus focused political discussions between people with different opinions[7, 8]. Because this freedom in social media can have a variety of negative effects for states and users, depending what their aims are, there have been many attempts to control it or take advantage of it in the past. Labeling posts as potential misinformation during the Coronavirus pandemic [13], monitoring or banning users that are suspected of being dangerous [17] or trying to influence elections through unconsciously shaping opinions of voters [15] are all applied methods by states and other people now in power. In authoritarian states the power even goes further. Active widespread monitoring and control is applied, to suppress opinions divergent from the state's goals and to enforce a certain type of thinking [17].

The model presented in this paper simulates the opinion dynamics of Reddit which is considered one of the least moderated popular networks. Nearly everything on Reddit is community regulated (bottom up), because the community decides which posts are seen by other members of the community. Our model analyses the effects of adding a moderator that tries to influence the user's opinions in his favor by the methods which can currently be observed in the real world (top down). We aimed to implement it as accurate as possible, using the open-source codebase of Reddit but as simplified as necessary to ensure proper performance. Further investigating the dynamics of social media platforms is of great importance, from analysing moderation techniques to discussing the optimal trade-off between personal freedom and avoiding polarization. Our model suggests that the attempt to control a social media network can enforce the desired behaviour if the right methods are used, but can also have opposite effects and stimulate extremist behaviour if users feel silenced or left out. In a complex system every intervention has to be thought out because unwanted side effects, as will be shown, can have huge consequences inside a society and especially inside a democracy, which can be strongly weakened by developments like polarisation or radicalisation[7, 10, 9]. These upcoming problems need answers to handle this fast development in our society right and thoughtfully.

2 Related Work

There has already been done a lot of research on the topic of opinion dynamics and especially opinion modelling, where each model follows its own approach of defining how agents can be defined and how they influence each other over time when they communicate. These elaborated models can be assigned to three different general approaches [5].

First there are models of assimilative social influence, which make the assumption that whenever two people interact, they influence each other towards reducing differences. Predominantly this was implemented by changing attitudes of agents towards the opinion of the other by some fraction of the "distance" (e.g. Abelson 1964, French 1956, Harary 1959). This assumption is based on psychological studies which had found in experiments of conformity that social influence reduces differences between people (Asch 1956). The main problem of this attempt is that it always leads to perfect consensus, when the graph representing a system is connected. This behavior of opinions is almost never empirically seen in a heterogeneous larger environment, like in a group, a country or in politics, and therefore has big limitations when trying to model real-life or social media dynamics.

The second category consists of models constraining themselves to similarity-biased influence only. In these, only sufficiently similar agents can influence each other's opinions by reducing opinion differences. There is scientific evidence which supports the phenomenon, commonly known as hemophilia. It can be summarized by the idea of "symbolic interactionism" (Stryker 1980), a theory stating that when new information is needed, people seek input preferably from sources/ people that they know share similar values to theirs[14]. These models can avoid consensus and mostly result in an emergence of opinion clusters. Thus performing better in modelling social media opinion dynamics, known for echo chambers and bubbles that form, this approach fits our research much better already[8].

The third category are models with repulsive influence. Their assumption is that individuals will not only assimilate or ignore each other. They will also distance themselves further from opinions they already disagree with (Rosenbaum 1984)[14]. The consequence of such an implementation is that often clusters form towards the extremes of the opinion spectrum[5]. This also fits opinion dynamics of social media and especially Reddit which is famous for having a lot of popular subreddits which are far left, far right or full of fake news.

Our model combines the second and third approach by having a similarity and dissimilarity biased influence. This was expected to result in a clustered opinion spectrum, ranging from clusters around the centre all the way to the extremes, which as explained comes close to the reality of Reddit. What differentiates our model from the vast majority of others is, that we prioritised the design of the environment our agents are interacting in. Because social media networks are of more and more relevance to our social life and our political opinion, and generally social activity increasingly shifts to the online world[12], where group and opinion dynamics are very different to personal contact, we not only focused on implementing an environment very close to Reddit. We also oriented the implemented moderation strategies around current approaches to moderation by social media companies. The most common strategies: Labeling of fake news[13] monitoring and banning users[17]. To our knowledge, there is currently no model that follows the approach of a Reddit-like simulation under the influence of moderation.
A big problem that is keeping opinion dynamics from bigger achievements is that "the field suffers from

a strong imbalance between proliferation of theoretical studies and a dearth of empirical work. More empirical work is needed testing and underpinning micro-level assumptions about social influence as well as macro-level predictions" [5]. In order to gain credibility and to establish consciousness in society and in politics about this field, further psychological and sociological research is needed to develop a foundation for these models which they can rely on.

3 Aims

The aim of this project is to understand the dynamics of opinion formation of a user base in a Reddit-like social network better on a macro level, especially when it is designed to influence the behaviour of this community in a certain way. This includes:

1. Designing a social network that is very close to Reddit, where agents can interact
2. Implementing moderation in this network which can use various methods to influence the community
3. Comparing various versions of moderation to gain information on the effects of the moderation methods

4 Our Model

Our model is implemented as a bipartite Graph $R = (U \cup S, E)$ where U is the set of users, S denotes the set of subreddits and an edge $e = (u, s)$ exists, iff user u has joined subreddit s. E is saved as an adjacency list, partitioned among the users. Since users can join and leave Subreddits on Reddit, so can our user objects. Therefore, our graph is dynamic.

4.1 Environment

The environment for the modeled agents is Reddit. Reddit is a social platform that consist of a number of Subreddits in which users can post and consume posts. In our model the set S of Subreddits is an array of Subreddit objects. Posts are defined to have an author, an opinion, the number of up- and downvotes and a timestamp of when it was created. We defined that Subreddits also have an opinion which is evaluated using the Subreddit's previous opinion and opinions from its most popular posts. This property of Subreddits is important for the selection process of users. The posts in a Subreddit are saved in two distinct data structures: a sorted-hot-queue and a new-stack, where hot is defined as a function depending on the logarithm of the difference of upvotes and downvotes as well as the age of the post, whereas new is automatically sorted by age. The hot function is the same as Reddit uses. The hot ranking is defined as:

$$\text{hot}(up, down, t_{current}, t_{creation}) = sign(up-down) \cdot \log_{10}(\max(|up-down|,\ 1)) + \frac{t_{creation} - t_{current}}{12} [2]$$

with

$$sign(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{if } x = 0 \\ -1, & \text{if } x < 0 \end{cases}$$

The network can be visualised as a graph which can be described as follows. Let $G = (V, E)$ be a bipartite graph, with $V = \{Users, Subreddits\}$ and E_{ij} = User$_i$ follows Subreddit$_j$. A post is defined as $P = (u_i, \alpha, u, d, v, t)$, where u_i is the user who created the post, α the post opinion, u the number of upvotes, d the number of downvotes, v the number of views and t the timestamp of when the post has been created. A Subreddit $S_j = (H, N, \gamma_j, F)$, where H and N are sets of posts ranked by Hot and New respectively, γ_j is the opinion of the Subreddit and F the amount of users. A User $U_i = (\beta_i, \beta_{online}, \beta_{create}, A_i)$, β_i being the user's opinion, β_{online} the probability of being online during any timestamp, β_{create} the probability of creating a post during any timestamp and A_i being the set of Subreddits followed by User i.

4.2 Agents

The agents, so called "Users" in this network are defined by an opinion, importance, β_{online}, β_{create}, β_{hot} and the list of Subreddits they follow. The opinion is modelled on a two dimensional continuous spectrum bounded by zero and one ($[0, 1]^2$). This allows users to adapt how strongly they feel about the two topics. The importance is a vector with two components, each one expressing how important a certain axis is to a user. Its euclidian norm is equal to the square root of 2, making it length-preserving on average.

A user has three bias values, namely β_{online}, β_{create}, $\beta_{hot} \in [0, 1]$, which define the probability of the user being online, creating a post and choosing to consume from the hot queue instead of the new queue.

Every user also features a variety of methods to interact with its environment. All together this is used to calculate whether a user agrees, ignores or disagrees with a certain post. To do so, we defined the measurement of how much a user agrees with a post as the L_2 norm between the user opinion and post opinion, each dimension weighted with the components of the importance vector. Each user is also defined by a dynamic list of Subreddits he follows. These are the Subreddits he can post to and consume content from. The user also has the ability to leave or join Subreddits, depending on if he is satisfied with the content he sees in the Subreddit.

4.3 Interactions

In every iteration every User decides on what to do. They have three options to probabilistically pick from:

1. Do nothing/ stay offline for an hour.

2. Create Post:
The User creates a Post-object whose bias is picked from a normal distribution around the Users bias. This Post is then posted to a randomly selected Subreddit. Obviously, Users will post only in Subreddits they joined.

3. Consume Posts:
The User randomly selects one of their Subreddits and looks at the first five Posts in the hot-queue or the new-stack. The User calculates their weighted bias difference, based on tolerance and bias. The interaction itself is split in two phases. Firstly, the User chooses to upvote, ignore or downvote the post, based on their weighted bias difference. Second, with more strict bias differences, Users can adapt their own bias to the one of the Post. If their weighted difference is less than 0.05,

adapted to the number of dimensions, the User will change their opinion, halving the distance to the post. If that difference is higher than 0.9 adjusted to the number of dimensions, the User will distance themselves by up to 0.01, opposite to the direction of their difference vector. If the User is already very extreme, however, it is more difficult to make him more extreme. If a User is unhappy with the post, they see, they can choose to leave the Subreddit. This can also happen on another occasion.

Another action a user can do is change Subreddits. This action is split into two components, joining and leaving. Whenever a user is online they are allowed to join a Subreddit that has similar opinions as the user with a certain probability. We define this probability as:

$$p(|A_i|, \text{SR-Cap}) = \frac{\text{SR-Cap} - |A_i|}{\text{SR-Cap}} \cdot 0.25$$

where $|A_i|$ is the number of Subreddits the user currently follows, and SR-Cap the number of Subreddits a user can maximally follow. For leaving a Subreddit we invert this probability. This term should be large if the user follows only few Subreddits and small if they follow many. This should incentivize the user to join some Subreddits but will not force them, as it could be that no Subreddits exist that have a similar opinion as them.

4.4 Moderation

The moderator agent stands above the entire network, being able to control posts and to make changes, such as changing a posts availability or excluding a person from the network. To make these decisions we introduced a zoning of the opinion spectrum. We define an opinion the moderator wishes the network to represent and introduce zones around this moderated opinion, each increasing the severity of the moderator's actions.

To do this, the moderator checks posts and calculates in which zone they lie. Stepping out of the "green" zone means that the post is considered slightly controversial from the moderators perspective. The Moderator marks the post so it does not get into the hot list and therefore controls how often this post will be seen and be able to influence people. This action can be seen similar to the labeling of posts done by social platforms in the past [13]. If a post lies in the next zone, the author is added to a blacklist, where he gets checked more regularly. This is equivalent to the monitoring done by China in WeChat [17]. When a post is detected that lies inside the most controversial zone, the moderator suspends the author for a day (24 Time steps, where each time step is 1h). A user notices these moderation attempts with a fixed probability and distances themselves 5% from the moderator's opinion as a repulsive reaction.

Some limitations still apply. The moderator can only look at a fixed amount of random posts every iteration and their blacklist has a fixed size as well, which should model that a moderator has a fixed amount of resources and can not check every single post and person in the network.

4.5 Parameter Choice

The network is initialized with a random distribution of opinions for Subreddits and users, connecting every user to Subreddits that match their opinions. For this we use the beta distribution, which models a populations opinion, where a certain opinion is shared more often. This models a demographic much better than a normal distribution and does not introduce issues with the boundary, that all opinions

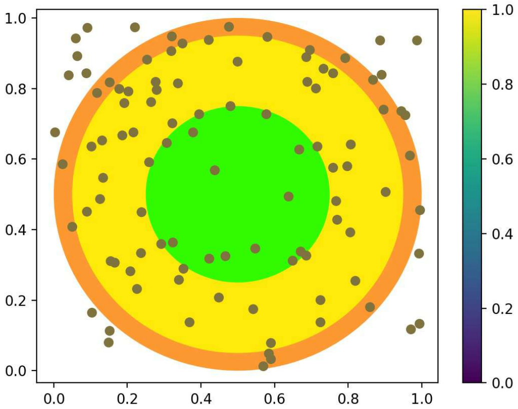

Figure 1: Moderation Zones overlaid over opinion spectrum

must lie within. The beta distribution is defined as:

$$f(x; a, b) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}[3]$$

Where the normalization, B, is the beta function,

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt$$

For user opinions we initialize one axis with the beta distribution with values $[3.1, 2.7]$ and the other with $[4.9, 5.2]$. These values are arbitrary and should reflect the population one models. The Subreddits are normal distributed in the beginning with values $[SR\_BIAS, 0.2]$, where we set SR_BIAS to $[0.5, 0.5]$. These opinions are only used to distribute users to Subreddits and will later be changed by the mean of popular posts on a Subreddit.

Whenever a user reads a post, an agreement function is called deciding how much this user agrees or disagrees with this post, depending on his opinion bias including the importance and the the post opinion:

$$\text{Agreement}(\beta_{User},\ \beta_{Post},\ \text{importance}) = \|(\beta_{User} - \beta_{Post}) \cdot \text{importance}\|$$

This function returns a value between 0 and $\sqrt{2}$. A user upvotes a post if the agreement function is smaller than 0.2 and downvotes a post if it is bigger than 0.8. The user gets influenced towards the post if the function returns a value smaller than 0.05 and away if it is larger than 0.9 (both adjusted to the number of dimensions). This choice of values makes the user much more relaxed about voting on a post and much more strict for the post influencing them. This models our intuition, that a post must be much closer and further to our opinion before it causes influence. If a User agrees with a post, their bias difference will half. If they disagree, they will distance themselves up to 0.01. This value is multiplied by the distance to the closest edge, making it harder for Users to take on extremely radical positions. After this calculation the opinion will always be constrained to lie inside $[0, 1]^2$.

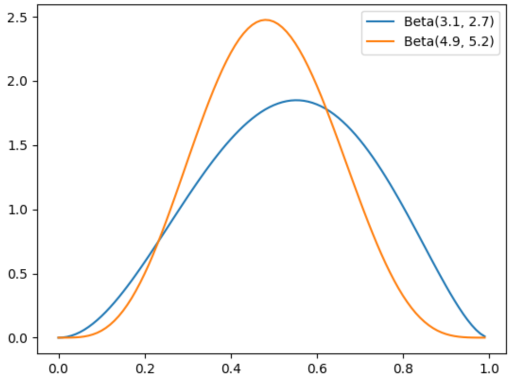

Figure 2: Beta Probability Density Function for Parameters [3.1, 2.7] and [4.9, 5.2]

5 Results

5.1 Simulation Runs

To evaluate the results, the model was simulated once without any moderation and four times with varying moderation capacities. In each run, we simulated 2000 Users on 112 Subreddits for a total of 480 rounds. Since there are around 50 million daily active users [6] on Reddit and 2.8 million Subreddits [11], the user-to-subreddit ratio is respected. In order to check our results, we ran the simulation multiple times, also for variable sizes and lengths. For the sake of simplicity we only show one result for each run. Table 1 below depicts the parameters used for the 4 respective runs:

Since the amount of posts per round was experimentally discovered to be around $P \approx \frac{U}{100} \cdot 1.6$, we deem the ability to check 100% of the posts as strong moderation capacity, and 12% as weak moderation capacity. Compared to Facebook, which manually reviews less than 1% of the photos and videos [4] uploaded, these figures seem comparatively large, however they are chosen so in order to account for automated reviewing as well. Moreover, we chose the tolerance zones for one moderator that allows broad opinions and for another who starts suppressing content close to the desired opinion already.

What matters as well is the moderator's ability to constantly and manually track a user present on the blacklist. Here, we classify the capability to monitor 10% of the users as strong, and the capability to monitor 1% as weak accordingly.

The moderator's bias value is fixed to $\beta_{\text{mod}} = [0.5, 0.5]$ for all simulations, since the scope of this paper is to analyse strategies to moderate extreme content from the moderators view and therefore the moderator considers itself as moderate.

Finally, in order to numerically measure the degree of extremism of the community and by this the

Simulations			
Moderation Type	Post Check Capacity	Blacklist Cap	Tolerance Zones
Type 0: None (Anarchy)	0	0	n/a
Type 1: Few Capacities & Weak Moderation	4	20	[0.35, 0.45, 0.5]
Type 2: Few Capacities & Harsh Moderation	4	20	[0.2, 0.3, 0.4]
Type 3: Strong Capacities & Weak Moderation	32	200	[0.35, 0.45, 0.5]
Type 4: Strong Capacities & Harsh Moderation	32	200	[0.2, 0.3, 0.4]

Table 1: Percentages of the User Segments for each Moderation Type

success rate of moderation, we count the users in the four zones we described before in the end and compare it to the start.

Extremity Measurement Parameters				
Segment	Centered	Slightly Nuanced	Strongly Nuanced	Extreme
$\|\beta_i - [0.5, 0.5]\|$	< 0.25	$[0.25, 0.4)$	$[0.4, 0.5)$	≥ 0.5

Table 2: Definition of the User Segments

5.2 Without Moderation (Type 0)

Starting off with every user's opinion beta-distributed on the political spectrum, letting them interact with each other unhindered resulted in clusters forming, which as stated before, is typical occurrence in social media networks. What can be observed when looking at the clusters, is that the biggest clusters are in the centre, while there are some smaller ones a bit further off. Thus, at all levels of the political spectrum, our agents reach consensuses with fellow similarly minded users, reproducing the behaviour of the similarity-biased influence models.

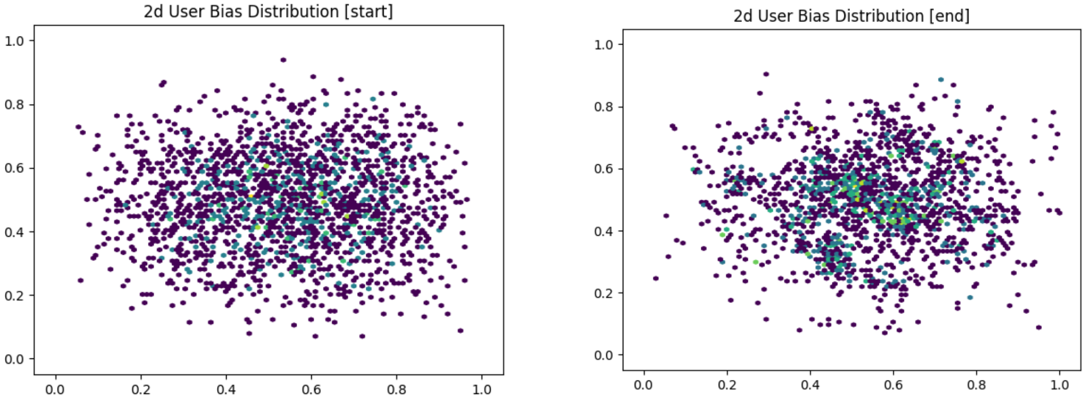

Figure 3: User Opinion Bias Distribution, Initially and after Simulation

Yet in our model, trends can change over time, which we can best measure by looking at the average (normalized) user bias. This is a strong indicator that the agents in our model managed to produce complex behaviour which cannot be reasoned about by simple explanation.

Another phenomenon that can be seen is that, while the number of extremist users increased slightly (from 0.35% to 0.55%), the number of neutral users grew by 7%. So although there hasn't been much change happening at the poles of extremity, there has been a small shift away from slightly differentiating from the neutral opinion towards either more extreme opinions or to more neutral opinions. Overall, the opinions of the entire network end up more divided, although not by much.

5.3 With Moderation (Types 1-4)

With moderation introduced, the problem of extremism hasn't vanished. No simulation, independent of its moderation type actually managed to decrease the total amount of extremist users (as defined in Section 6.1). However, the differences between the outcomes when choosing other moderation types are clear and significant.

Firstly, a Type 1 moderator actually manages to draw more users to the center, as the number of neutral users grows by 18%. Mostly all of these users come from the slightly and strongly nuanced users, which taken together decrease roughly by the same absolute amount. So aside the fact that the number of extremist users grew by 50% relative to the start, this moderation actually managed to "moderate" a lot of users as can be seen in Figure 4 below. Type 2 Moderation made the extremist population double, however compared to Type 1 moderation, the effects were similar, with a lot of slightly nuanced users becoming neutral, but the changes were oddly less radical as in Type 1 Moderation, despite enforcing harsher moderation.

Type 3 and 4 moderation eventually showed that a moderators capabilities mattered just as much

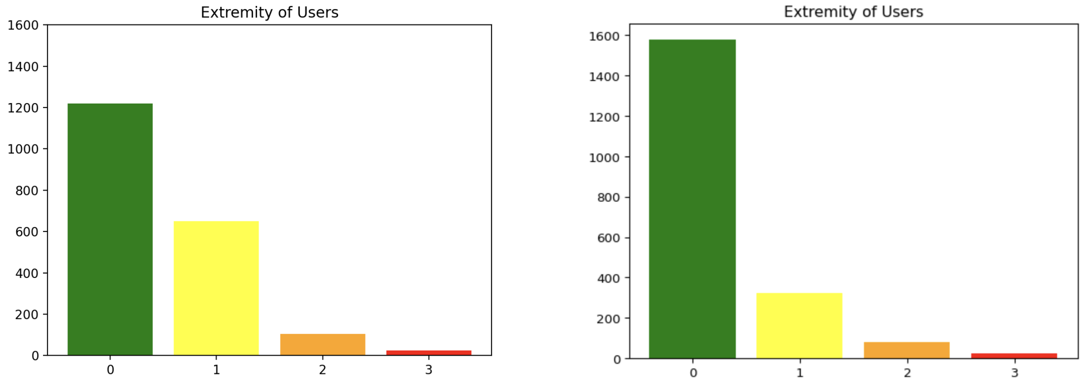

Figure 4: Population of the User Segments before and after Type 1 simulation

in the development of extremist user formation as the harshness of the moderators regime. Both moderation types brought down, contrary to moderation types 0, 1 and 2, the total number of neutral users and managed to decrease the number of slightly nuanced users as well. Most notably, Type 4 moderation, a harsh and well-equipped approach to moderating, reduced the number of neutral users from 60% all the way down to 35%. Of equal if not greater significance, extremism grew 9-fold with Type 3 moderation and 25-fold under Type 4 moderation. And this despite that the success of extremist posts isn't greatly affected by the moderation type and the huge number of successful posts in the neutral range under Type 4 moderation. This means that most of the extremism doesn't stem from gradual radicalization by consuming posts, but mostly from the disagreement with the moderator itself.

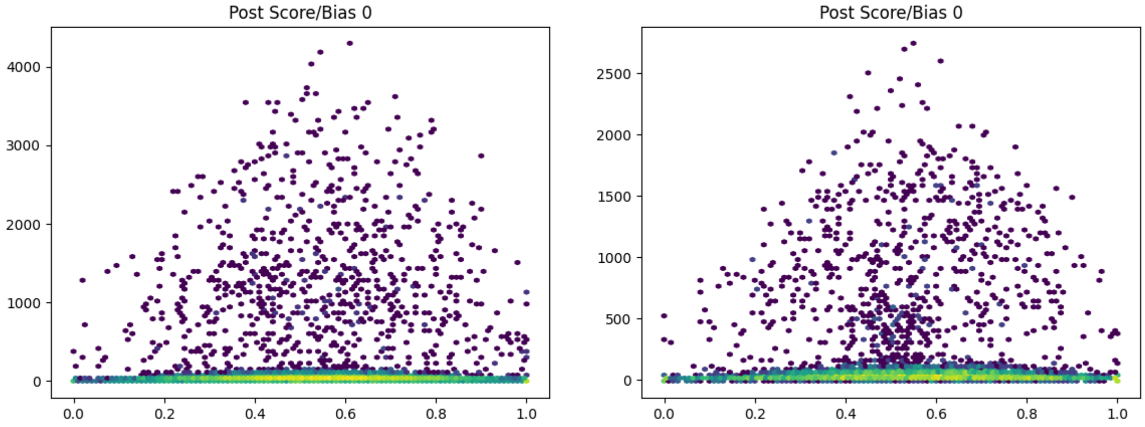

Figure 5: Success of Posts sorted by their first Opinion Bias value for Type 1 and Type 4 Moderation

Results										
	Type 0 (start)	Type 0 (end)	Type 1 (start)	Type 1 (end)	Type 2 (start)	Type 2 (end)	Type 3 (start)	Type 3 (end)	Type 4 (start)	Type 4 (end)
Neutral	61.2	68.7	61.6	78.75	61.55	66.85	60.75	57.8	60.5	34.8
Slightly Nuanced	33.15	27.7	32.4	16.05	32.8	26.3	32.85	20.05	33	25.75
Strongly Nuanced	5.3	3.05	5.25	4.05	5.15	5.85	5.7	12.85	5.75	20.8
Extreme	0.35	0.55	0.7	1.15	0.5	1	1.15	9.3	0.75	18.65

Table 3: Proportion of the total population represented by each user segment (in percent)

6 Discussion

6.1 Our Model

Our model clearly demonstrates the adverse effects strong (and absolute) moderation can have on the polarization of opinions among the Reddit user base and how less moderation actually can result in a less divisive opinion spectrum. When choosing the best moderation approach, one must first define what well-working moderation actually means. For some, well-working moderation may entail bringing the most amount of users to the center of the opinion spectrum (or to a specific point), whereas for others it is all about keeping a healthy balance in the user segmentation as defined in Section 6.1. With our model, Type 1 Moderation works exceptionally well in bringing the users to the centre, whereas no moderation and Type 2 moderation score best in minimizing the variance from the start and the end share of the respective segments.

6.2 Future Research

An opinion dynamics simulation in a user graph simplifying Reddit has been adequately achieved by our model. However, there are still numerous extensions one could consider to come to more interesting and realistic results. A crucial property we excluded from our model for simplicity's sake is the comment function under posts. Users mostly don't influence themselves by consuming posts only, but also by engaging in debates in the comments section of the post. [1] A model including the property thus may be able to model the activity on Reddit more accurately. What one could also consider is giving the opinion space more dimensions, and initialise the values of the user opinion biases from real-life distributions, for example by modelling answers to survey questions on one dimension each. The model is already built to support that - with approximately similar results.

The actions and interactions in our model are mostly decided by probabilistic formulas. Such probabilistic decision-making may be able to reproduce complex social behaviour, but not necessarily the outcome of a user base that can have tremendous amounts of differing reactions to introduced moderation procedures, given the numerous different contexts they are implemented in. For example a user may continue posting extremist content if the consequences are low, contrary to a user who may face harsh real-life punishment like a hefty fine or a prison sentence. Thus substituting probabilistic decision-making through the introduction of intelligent behaviour, e.g. by giving every user a simple neural network that determines his actions, one may be able to model out more intricate behaviours of social media users when reacting to moderation measures, since it would be easy to add more contextual parameters.

The scope of our paper was limited to the analysis of extremism in moderated Reddit-like networks, which was measured by taking the distance of a user's bias value from the centre. This model however, would also be able to analyse moderation techniques not designed to limit extremism, but to curb opinions that are not conforming to the values of an authority. Such results may be of great interest, especially given that countries like China, Myanmar, Turkmenistan and Russia are actively censoring content on the internet and may help us understand the opinion dynamics better of the people living in censorship-implementing countries.

6.3 Conclusion

Overall, this model finds that weak global moderation on Reddit can have slight positive impacts on minimizing polarization, but adverse effects increase substantially the stronger and stricter moderation is implemented. We were able to show this by evaluating the results of 4 different moderation techniques, which could be ultimately compared with the result in a moderation-free Reddit network. As numerous social media companies continued to strengthen their moderation approaches over the years and since Reddit stands amongst the most popular social media sites and plays a huge role in opinion influencing around the world, our model can give a simplistic overview of the dangers global moderation can have on Reddit and its users.

References

[1] *Engaging Reddit: Commenting and The Importance of Proactive Communication.* https://thebetterwebmovement.com/engaging-reddit-commenting/. Accessed: 2021-12-20.

[2] *How Reddit ranking algorithms work.* https://medium.com/hacking-and-gonzo/how-reddit-ranking-algorithms-work-ef111e33d0d9. Accessed: 2021-12-19.

[3] *Numpy Beta Distribution API reference.* https://numpy.org/doc/stable/reference/random/generated/numpy.random.beta.html. Accessed: 2021-12-19.

[4] *The Real Challenge in Content Moderation.* https://visua.com/challenge-in-content-moderation. Accessed: 2021-12-15.

[5] Thomas Feliciania Edmund Chattoe-Brownb Guillaume Deffuantc Sylvie Huetc Andreas Flachea, Michael Mäsa and Jan Lorenz. *Models of Social Influence: Towards the Next Frontiers* . https://www.jasss.org/20/4/2.html. Accessed: 2021-12-19.

[6] David Curry. *Reddit Revenue and Usage Statistics (2022).* https://www.businessofapps.com/data/reddit-statistics/. Accessed: 2021-12-19.

[7] Maeve Duggan and Aaron Smith. *The tone of social media discussions around politics.* https://www.pewresearch.org/internet/2016/10/25/the-tone-of-social-media-discussions-around-politics/. Accessed: 2021-12-19.

[8] ECPS. *Echo Chamber.* https://www.populismstudies.org/Vocabulary/echo-chamber/. Accessed: 2021-12-19.

[9] Thomas B. Edsal. *We're Staring at Our Phones, Full of Rage for 'the Other Side'.* https://www.nytimes.com/2022/06/15/opinion/social-media-polarization-democracy.html. Accessed: 2021-12-19.

[10] Michael Jensen. *The Use of Social Media by United States Extremists.* https://www.start.umd.edu/pubs/START_PIRUS_UseOfSocialMediaByUSExtremists_ResearchBrief_July2018.pdf. Accessed: 2021-12-19.

[11] Ying Lin. *10 Reddit statistics every marketer should know in 2023 [Infographic]* . https://www.oberlo.com/blog/reddit-statistics. Accessed: 2021-12-15.

[12] Ben Quinn. *Social network users have twice as many friends online as in real life.* https://www.theguardian.com/media/2011/may/09/social-network-users-friends-online. Accessed: 2021-12-19.

[13] Tianyi Wang Timothy R. Tangherlini Vwani Roychowdhury Shadi Shahsavari, Pavan Holur. *Conspiracy in the time of corona: automatic detection of emerging COVID-19 conspiracy theories in social media and the news.* https://link.springer.com/article/10.1007/s42001-020-00086-5. Accessed: 2021-12-19.

[14] Daniel G. Stefano Balietti, Lise Getoor and Duncan J. Watts. *Reducing opinion polarization: Effects of exposure to similar people with differing political views.* https://www.pnas.org/doi/abs/10.1073/pnas.2112552118. Accessed: 2021-12-19.

[15] Wikipedia. *Cambridge Analytica.* https://de.wikipedia.org/wiki/Cambridge_Analytica. Accessed: 2021-12-19.

[16] Wikipedia. *Cencorship, Internet.* https://en.wikipedia.org/wiki/Censorship#Internet. Accessed: 2021-12-19.

[17] Jing Yang. *WeChat Becomes a Powerful Surveillance Tool Everywhere in China.* https://www.wsj.com/articles/wechat-becomes-a-powerful-surveillance-tool-everywhere-in-china-11608633003. Accessed: 2021-12-19.