A Study of Facebook's Censorship Policies

Rubab Zahra Sarfraz
16030024@lums.edu.pk
Lahore University of Management Sciences
Lahore, Pakistan

Syeda Fatima Naqvi
16030007@lums.edu.pk
Lahore University of Management Sciences
Lahore, Pakistan

ABSTRACT

Censorship is becoming an increasingly interesting topic today. With its impact on the users and the kind of response it receives, it has become important to investigate what makes the social networking websites take these steps. We have carried out a study on Facebook's censorship policies; what Facebook declares in its community standards and how it ensures them. We have looked at events of uprisings from users, where Facebook bothered and where it did not. We have also carried out experiments on Facebook to understand the exact process of moderation. Our results show that Facebook takes censorship seriously but needs to be reminded by the public in order to draw its attention to certain sensitive areas.

1 INTRODUCTION

Over the past few years, social networking websites such as Facebook and Twitter have become extremely popular. What started off merely as a platform for interacting and catching up with friends has evolved into much more; a forum where people discuss and debate over the current global situation. To ensure that nothing inappropriate is said or shared, these social networks have set up some moderation policies. If anyone is found or is reported to have violated these, their post is censored and/or their account is blocked. However, oft times we have come across incidents where Facebook itself is reported to have violated its moderation policies by not standing by its word. Many journalists have written about events where content was unduly censored or deliberately ignored. Some of the most popular recent events include the Kashmir issue, where posts in favour of Kashmiris were censored; the 'Napalm Girl' issue [1], where a picture from the Vietnam War posted by a journalist was removed for showing nudity; the video of a policeman murdering a black woman in the U.S. being taken down from Instagram, and many others.

In light of how commonly available social media has become, it is important to realize that it vastly impacts the perception of its billions of users [7]. The owners of these social networks have the power to mould public opinion by what they choose to show and what they choose to hide. Censorship is, therefore, a sensitive issue and needs to be dealt with vigilance and care. Whether those in authority follow the rules set by themselves and remain unbiased is an important question that needs to be answered.

We have targeted Facebook as the main focus of our study as it is by far the most widely used social networking website in the world. A recent survey conducted in March 2017 shows that Facebook has a total of 1.94 billion users followed by YouTube with 1 billion visitors (which is only about half the number) [3]. In addition to that, Facebook has a defined set of censorship policies and a team of around 4500 content moderators. Facebook also claims to be devoted towards improving its censorship policies and the experience of its users. However, there have been multiple incidents (some of which are mentioned above) where Facebook's censorship policies have been questioned and brought into the limelight by its users. Many of these incidents have also forced officials from Facebook to speak up and defend their actions. For instance, a few years back a Muslim from Pakistan fought to have Mark Zuckerberg sentenced to death because Facebook refused to ban content about Prophet Muhammad that offended him. In response to this, Mark wrote a post defending his views on the matter by calling that piece of content "freedom of speech". In other cases, Facebook has been made to reconsider its actions and unblock a removed post or remove previously allowed content. For example, in the case of the Vietnam War photo, the rebuke from Norway's largest newspaper and the prime minister himself forced Facebook to restore this photograph. At another time, a women's rights protection group protested against the gender-based hate speech on Facebook and was able to get Facebook's attention on the matter. Most of the photos they reported were removed.

In our study we have tried to gain insight of Facebook's censorship mechanism and whether there exist any limitations or biases. For this purpose, we first carried out a detailed study of their community standards and events where Facebook is said to have allowed violations of its policies. This is mentioned in section 2 of the paper. In section 3, we go on to explain how Facebook implements its censorship policies. Section 4 covers the experiments we carried out in order to detailed view of Facebook's censorship mechanism and the deductions we have made from the results of our experiments. Section 5 and 6 cover the future work and conclusion respectively.

2 FUNDAMENTAL GOALS

The major goals of this study revolve around deeply understanding the censorship policies of Facebook and finding how consistent Facebook is in their implementation. Following are the questions that we seek to answer through our work:

(1) Is the content censorship by Facebook in accordance with its moderation policies?
(2) If that is not the case then what could be the possible factors governing this difference? Is there any bias towards any religion, race or country?
(3) What is the public reaction towards this censorship?

To answer these questions, we started off by studying Facebook's community standards in detail. These can be viewed at [2].

2.1 Community Standards

Facebook has set out certain rules and regulations to maintain a positive and friendly environment on the social network which it calls its Community Standards. These standards are to be followed by all Facebook users. If any user feels that someone or something

on Facebook goes against these standards, they can report it to Facebook. Facebook claims that it has set out these policies because it wants its users to feel safe when using it. On the other hand, it also warns that due to the diversity of the Facebook community, something that might seem disturbing or wrong to one group may not violate their community standards as it may be alright for others. This is where we need to look; does Facebook draw this line fairly or do we see a bias towards certain communities?

2.2 Public Opinion

To answer the question above, we researched about different events which had caused Facebook users to complain about Facebook's censorship policies. We found that in most cases people were unhappy and felt that Facebook is biased. The list of groups offended by these policies includes LGBTQ groups and individuals, artists, museums and galleries, Europeans, cannabis advocates, journalists, indigenous groups, sexual health organizations, plus-sized women, mothers, and women in general [4]. Some concrete examples that we found are mentioned here.

One of the most recent examples is that of the Kashmir issue where posts of influential people speaking against the brutality of the Indian army in Kashmir were found to be taken down and their accounts blocked. Huma Dar, a scholar at UC Berkeley and a political and social activist, wrote that her account was permanently blocked and all her data was lost beyond recovery for posting in favour of the Kashmiri Muslims [6]. An Indian professor at the University of Westminster, Dibyesh Anand, also had his post removed for speaking against the Indian army on the same issue. However, his account was soon recovered along with apology messages from Facebook. Figure 1 shows his post that was removed. Here we see a contrast in the reactions to both posts which shared similar context.

In another case we saw that Facebook's community standard statement that "Facebook removes hate speech, which includes content that directly attacks people based on their sex, gender or gender identity" was clearly violated but Facebook was remaining silent on the matter. There was a volume of content containing gender-based hate speech on Facebook which was repeatedly being reported by members of the Women, Action the Media team but to triggered no response from Facebook. After over 60,000 tweets and 5000 emails, Facebook finally took action and removed the content full of violence and hate speech against women.

3 HOW FACEBOOK CENSORS

With the recent release of Facebook's quarterly results of 2017, the scale of Facebook's reach is quite evident and it is increasing day by day. Following are some interesting statistics about Facebook [3].

(1) It has more than 1.9 billion people, including almost 1.3 billion people active every day
(2) Every 60 seconds on Facebook: 510,000 comments are posted, 293,000 statuses are updated, and 136,000 pho- tos are uploaded.
(3) Five new profiles are created every second.
(4) There are 83 million fake profiles.
(5) Photo uploads total 300 million per day.
(6) Average time spent per Facebook visit is 20 minutes.

Figure 1: An example of a post censored by Facebook.

Given this amount of content being generated on seconds of timescale, Facebook needs to have a robust mechanism to moderate its content in order to provide a healthy virtual environment to its users.

3.1 Facebook's Content Moderation Approach

Although Facebook revealed its high level view of handling reporting process back in 2012, it still isn't very open about its content moderation methodology to public. Some approximations can be made by connecting different bits and pieces from Facebook's official sources and get a useful insight into its moderation process. If we look at the scale of data that Facebook has to moderate, relying on just algorithms or automated scripts can prove to be a challenging task. Although, several advances have been made in the field of artificial intelligence in recent years, such as Facebook is currently using AI to report offensive/harmful pictures before they hit the public audience and it has reported more such pictures than humans [5] but it still has long way to go before taking over the complete responsibility of content moderation because user-generated videos are too complicated for computer vision to handle and con- tent requires a considerable context. Also, relying on solely human-based moderation would require a lot of man power and time. By the time, content moderators make a decision, damage would already been done. Hence, Facebook uses a middle ground and deal with censorship by creating a system consisting of both technology as well as humans. All the content that gets posted first goes through the algorithm which processes it. Our current discoveries tell us that it processes links and images and after passing them through relevant checks, it allows or disallows them. Once the content is successfully published, any user can report it and it gets submitted to Facebook's content moderation team for review. It has been known that Facebook's content moderation team consists of 4500 moderators and it also outsources its work to companies across the globe such as CrowdFlower and in countries like Phillipines, Germany, India, etc. [5].

Table 1: A list of categories covered in Facebook's Community Standards.

No.	Categories of Community Standards
1	Direct Threats
2	Self-Injury
3	Dangerous Organizations
4	Bullying and Harassment
5	Attacks on Public Figures
6	Criminal Activity
7	Sexual Violence and Exploitation
8	Regulated Goods
9	Nudity
10	Hate Speech
11	Violence and Graphic Content

3.2 Reporting Flow

Figure 3. depicts different options given by Facebook when reporting a post, group, page and a person. As it can be seen, Facebook broadly classifies content into the categories mentioned in their community standards. An interesting thing to note here is that every type of content does not necessarily gets submitted for review by users. That's probably because to reduce work on moderators end. Only, if you choose I think it should not be on Facebook and spam, then you get to submit that piece of content for review. Status of all the reports submitted by a user can be seen in support inbox of Facebook.

Figure 2. shows a high level view of how content gets moderated. Firstly, when a user finds something reportable, he or she reports it by selecting appropriate option from the menu and submits it for review. The post is then forwarded to the system that deals with distribution of content which forwards it to content moderators. Upon receiving some content for review, the content moderator decides to allow it, block it or escalate it to headquarters (if it is outsourced) according to the polices and rules provided by Facebook to them. If the moderator decides to allow it, it would be informed to the reporter that that specific content did not violate any community standard. If it is to be escalated, it is forwarded to US-based team for help in cultural context and California-based team for handling threats and dangerous situations. If the moderator decides to remove it, the reported user gets told about the action taken for that purpose. The person who posted the content gets contacted in this case and based on the severity of the action, gets penalized, for instance, that content could get removed, Facebook could block the user from posting anything in future etc. Facebook also collects active feedback on both the ends i.e. reporter as well as the person whose content gets reported, both the parties are provided with the feedback forms that how satisfied are they with their experience which they can answer according to their experiences.

4 EXPERIMENTAL ANALYSIS

We started out our experimentation by using Facebook's test accounts because they are exempted from Facebook's spam and fake account detection systems. But they have certain limitations:

(1) They cannot interact with real accounts hence we would not be able to post on other people's walls and comment on their posts etc.
(2) They have restricted view of Facebook with no search bar hence we cannot access any page or profile.
(3) Since we can't comment/post on pages, our content's reach would be limited.

Hence, test accounts did not meet our requirements so we went on to create fake accounts.

4.1 Probing Facebook from Fake Accounts

We created two fake accounts, a male and a female one. Both represent different cultures and religions. We also formed a team of people on Facebook who reported the content that we posted through these accounts. We posted content that have already been censored by Facebook in the past as well as new content so as to see if their censorship has some dependence over time. We created a database in which we stored the following metrics:

(1) `Category:` The type of the content according to moderation policies.
(2) `Reported_at:` The timestamp at which the content was reported.
(3) `Account:` The fake account that posted the content.
(4) `Response:` The response by Facebook's content moderation team.
(5) `Response_at:` The time at which review report was sent back.

4.2 Types of Categories

Initially, we have targeted some of the categories from moderation policies by Facebook and posted content that fell under these specific categories. We chose these because of their commonality and also they have instilled some controversies in recent times. These categories are:

(1) **Gender-based hate speech:** This category included posting pictures and content which were explicitly against women. These included sexual assault, physical abuse, domestic violence as well as sadist jokes against women.
(2) **Nudity and sexually explicit content:** This type included content that was physically explicit. Both types of content were posted, that should have been removed as well as that should not have been.
(3) **Graphic violence:** This covered images containing graphically offensive content such as beheadings, child abuse and so on.

4.3 Results

Table 2. shows the results of our experimentation. Average response time is the time taken by a post to get reviewed. Out of the total content posted, 76% was such that it violated Facebook's community standards. By positive reviews, we mean the posts that should have been removed and actually were. These posts made 52.6% of total. Rest of the percentage was of negative reviews, those posts which should have been removed but were not. These results show that although Facebook is trying to be fair in implementing its policies,

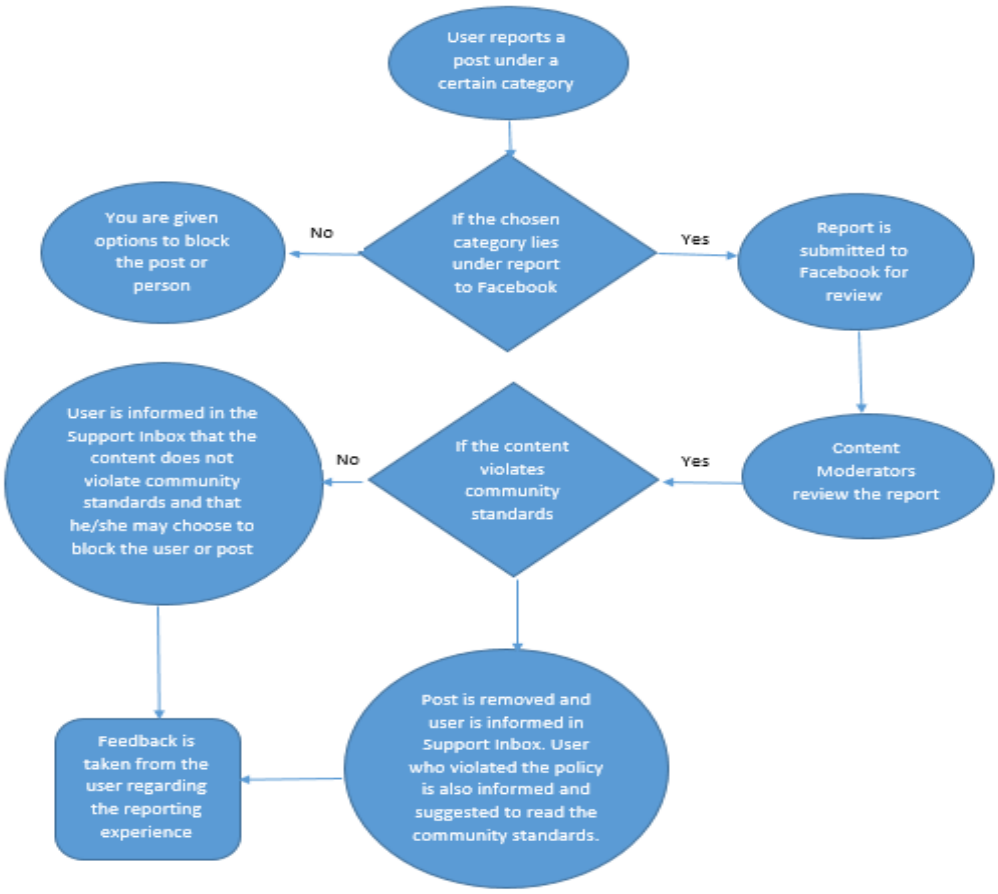

Figure 2: Flow diagram of Facebook's content reporting process.

it's not entirely successful in doing so. In our results, we have also seen that all those posts which should not have been removed were not. This shows that Facebook is not too strict in implementing its policies.

5 DEDUCTIONS

Through our study of various censorship events and experiments we have made certain deductions. Currently, Face- book's content moderation algorithm only moderates images and links. Spam links are blocked in real time whereas images are flagged and later reviewed by human content moderators. If a user's content gets removed their account is kept under observation depending on the gravity of the situation. Such users may face some restrictions on posting content etc. In case of repetitive violation of community standards, sometimes users are kept from accessing their account for a certain period or their accounts may even be permanently blocked. Some categories have more fine-grained policies than others depending on the sensitivity of the topic. Certain policies, regarding gender-based hate speech for example, have been refined due to protests by the users. Since, not all policies are fine grained there is a chance that in such scenarios the subjectivity of the content moderators' creeps in. We have observed that the reports received after review do not contain enough details about violation of policies. It would be helpful if they mention the exact clause that have been violated by the user. Another observation is that reporting a person, page or a group means Facebook would be verifying their authenticity only and not the content they have posted. We feel that Facebook's huge scale is a major challenge for them. The scalability of their content moderation techniques needs to be focused on. Recently, Mark Zuckerberg stated that they are adding 3000 more members to the content moderation team to deal with moderation more effectively and prevent mishaps such as suicide, murder, and other such criminal activities. Following the current trend of rapid advancement in AI, certain menial tasks can be automated thus reducing the subjectivity introduced by human content moderators.

6 FUTURE WORK

We feel that an important consideration at this stage will be the direct, first-hand user opinion. For this purpose, surveys can be conducted asking Facebook users what they feel about the censorship policies and how they affect them. These surveys should, ideally, be targeted at as wide a user range as possible. The more diverse the set of users is, the more insightful will be the results. Moreover, we

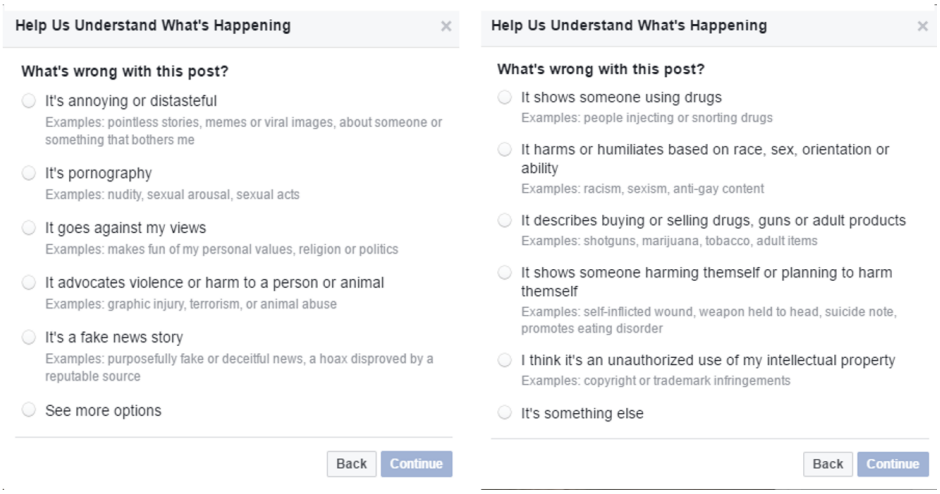

Figure 3: Facebook's options for reporting a post.

Table 2: Empirical Results from Experiments. It shows that Facebook still has a long way of achieving more than 50% accuracy in each category.

Statistic	Value
Average Response Time	2.38 hours
Posts violating community standards	76%
Posts not violating community standards	24%
Percentage of Positive Reviews	52.6%
Percentage of Negative Reviews	47.4%

could expand our horizon by experimenting with more categories of content such as illegal activities, video content etc.

7 CONCLUSION

After reading through a lot of controversies stirred against Facebook regarding censorship, we have designed and con- ducted experiments to verify and test certain beliefs about Facebook's approach towards its content moderation. Al- though Facebook's willingness to approach this problem is commendable, as many social media platforms do not bother taking responsibility of the content posted to them, our comprehensive experimentation shows that the dissatisfaction among Facebook's users is valid and it should take further steps to improve its censorship policies as well as their implementation in order to reach at a point that Facebook envisions [8].

REFERENCES

[1] V. Doshi. 2016. *Facebook under fire for 'censoring' Kashmir-related posts and accounts.* https://qz.com/719905/a-complete-guide-to-all-the-things-facebook-censors-hate-most/ The Guardian.
[2] Facebook. [n.d.]. *Facebook, Community Standards.* Technical Report. Facebook.
[3] Facebook. 2017. Facebook-Q1-2017-Earnings. *https://investor.fb.com/investor-events/event-details/ 2017/Facebook-Q1-2017-Earnings/* (2017).
[4] Y. Jillian. 2016. *A complete guide to all the things Facebook censors hate most.* https://qz.com/719905/a-complete-guide-to-all-the-things-facebook-censors-hate-most/ Quartz.
[5] C. Josh. 2016. *Terminating Abuse.* https://techcrunch.com/2016/05/31/terminating-abuse/ Techcrunch.
[6] Women Action Media. 2016. *Examples of Gender-Based Hate Speech on Facebook.* http://www.womenactionmedia.org/examples-of-gender-based-hate-speech-on-facebook/
[7] D. Morrison. 2014. *Trends and Applications in Knowledge Discovery and Data Mining, chapter Toward Automatic Censorship Detection in Microblogs.* Springer International Publishing.
[8] M. Zuckerberg. 2017. *Building Global Community.* https://www.facebook.com/notes/mark-zuckerberg/building-global-community/10154544292806634/ Facebook.