

B. Des. Project II

Whatsapp Group Moderation Featureset

Rishabh Jain

16U130015

Guided by

Prof. Venkatesh Rajamanickam



IDC School of Design
अभिकल्प विद्यालय

Declaration

I declare that this written project submission represents my own ideas and work in my own words and if any idea or work which has been included, I have adequately cited and referenced the original source.

I also declare that I have adhered to all the principles of academic honesty and integrity and have not misinterpreted, fabricated or falsified any idea/ data/ fact source in my submission.

I understand that any violation of the above will be cause for disciplinary action by the institute and can also evoke panel action from the sources.

Signature:

Name: Rishabh Jain

Roll No: 16U130015

Date:

Approval Sheet

The BDes Design Project II titled “Whatsapp Group Moderation Featureset” by Rishabh Jain, Roll Number 16U130015 is approved, in partial fulfillment of the Bachelor in Design Degree at the IDC School of Design, Indian Institute of Technology Bombay.

Program Co-ordinator:

Chairperson:

Table of Contents

A Note on the Topic	2		
Acknowledgement	2		
Introduction	3		
1.1 Design Brief	4		
1.2 Goals	4		
1.3 Non-goals	4		
1.4 Scope	4		
1.5 Challenges	4		
1.6 Video Presentation	5		
Secondary Research	5		
2.1 Platforms: Content focussed	5		
2.1.1 Reddit	5		
2.1.2 Facebook Groups	7		
2.2 Messaging Applications: Communication focussed	8		
2.2.1 Discord	8		
		2.2.2 Telegram	9
		2.3 Insights from BBCs Duty, Identity, Credibility Report	10
		2.3.1 Content norms: Summary	10
		2.3.2 Off Topic Conversations	11
		User Stories	12
		3.1 Cricket Group	12
		3.2 Class Group	12
		3.3 Online Learning Group	13
		Proposed Changes	14
		4.1 Current v/s Proposed states of Whatsapp	16
		4.1.1 Current State	16
		4.1.2 Proposed State	17
		4.1.3 Free Speech Implications	19
		4.1.4 Non-Proposal	20
		4.2 Implementation	21
		4.2.1 Participant Tags	21

4.2.2 Message Status Prompts, Group Status Prompts	22
4.2.3 Slow-Mode/Conversation Mode	23
4.2.4 Group Settings	23
4.2.5 Blacklist	23
4.2.6 Report	26
4.2.7 Review (Moderation)	27

Conclusion	29
Reflection	30
Bibliography	31

A Note on the Topic

This journey started with different ideas than the ones presented here. I picked the domain of Fake news and wished to address it with a game at the start of the semester. It was only after diving deep into it that I realised that I was way out of my depth there. I got lost in its philosophical foundations that dealt with things like truthiness, epistemic responsibility & post-truth. If this was the only place I got lost in, that might have been fine, but my vision kept getting murkier with added layers of complexity. There are systemic reasons for platforms prioritising engagement over signals of truth. Finding the truth is hard for the machines and both hard & expensive for humans. Teaching people how to find the truth online is an added challenge. Media literacy techniques optimised for the web are few, far between and limited in their results. In addition, this domain is directly affected by the flaws in human reasoning like biases and motivated reasoning.

As a design student, I found it impossible to define a design brief and find a direction. It did not help that I kept questioning my own assumptions rather than acknowledging them and building on them.

After a series of failed attempts at connected but distinct approaches, I have finally fixed on the following brief. It is a necessary but insufficient precondition to a fake news intervention.

Over the span of these past months, I have reached the conclusion that fake news may not be a design problem at present. I aim to

articulate and write a piece of literature arguing the same but it was deemed out of scope for this particular project.

Acknowledgement

I would like to express my sincere gratitude towards my guide Prof. Venkat, for helping me deal with the turmoil in this project. I would also like to thank Prof. Dalvi for his inputs.

My friends, Tarun, Advait, Malay, Gyan & Rishi have been invaluable in providing support, direction, and insight into my own thinking through many extended discussions. They were also much needed sources of encouragement and inspiration in these troubling times.

Towards the end of the fake news saga, Paulanthony George's work mapping web historiography of news, misinformation and disinformation helped me to better articulate my own reasons for not pursuing it further.

Last but not least, my family, for making efforts to provide the conditions needed for work during the Covid-19 pandemic.

1. Introduction

According to Whatsapp's website(Whatsapp, n.d.), it is mainly used by people to stay in touch with family and friends. In contrast to that, Whatsapp is now a much broader platform with people using it not only to connect with friends and family but also to broaden connection with different communities through groups. These groups have implicit and explicit rules about what content is acceptable on them. Currently, Whatsapp offers limited to no tools to manage these communities. In India, these groups can be based on workplace connections(formal office groups), professional associations(eg CAs in Thane), around common interests(cricket, board games, books etc), around events(Conference, birthday party, marriage planning), around Socio-Political identities(Modi Supporters, Kanyakubja Samaaj) etc. This is unlike other online communities where the admins have access to advanced forms of moderation. The justification for this would be that these platforms usually deal with a large number of unknown people with varying goals, some of which might be nefarious and hence it is necessary to be able to regulate it. Whatsapp is also at a stage in its adoption and usage

where these tools are now needed. It is not a friends and family only communication platform anymore.

Whatsapp has also been the centre of multiple controversies in the past regarding the sharing of information that has led to horrific outcomes like mob lynchings and mob violence. These messages are typically believed to spread through groups because they offer a larger spreading surface. Whatsapp has been approached by various governments and concerned citizens to take action against such misuse of the app. Whatsapp has reduced the number of possible shares at a time for highly forwarded messages, first to five and now to one in the wake of the coronavirus pandemic(Singh, 2020).

According to Whatsapp these steps have aided in reducing the spread of viral information. The steps taken are in the right direction but not enough to tackle the problem. There have been some incidences(Sikdar, 2016; Roy, 2017; YS, 2017) and some rumors(FE, 2020) about the government prosecuting group admins for allowing/facilitating the spread of hate content but little has been done in this regard. This is a two fold problem, on one hand any regulatory body is not privy to the conversations within a closed off whatsapp group and on the other is that the admins have little control on what people post in the groups and can shrug off responsibility. In the future, even if it is a legal responsibility, they simply do not have

the tools needed for the job. This project aims to design such moderation tools. This can curb the argument about the inability to monitor group content and hopefully lead to an increase in accountability of group admins on the kind of content they allow to flourish in the communities they manage.

1.1 Design Brief

How might we empower Whatsapp Group admins to moderate the content on the groups they manage?

1.2 Goals

- Enable admins to regulate the content within reasonable boundaries of group norm
- Keep admin activity visible to the rest of the group

1.3 Non-goals

- Educate admins/members about quality control
- Curb Fake news/hate speech explicitly
- Introduce Algorithms/Third party moderation

1.4 Scope

- Studying the other tools that offer similar functionality
- Study the Whatsapp Design language
- Create High fidelity mockups for the concepts

1.5 Challenges

Whatsapp offers a unique problem because of its privacy focused end to end encryption. Encryption is a double edged sword in the pursuit of monitoring/controlling content that flows through a platform. While it is essential that services use encryption to protect user privacy, it makes it next to impossible for researchers/platforms to be able to pin down consistent sources of objectionable content. The inability to access the content is detrimental to scalable algorithmic/human moderation of any kind by the platform itself, which are commonly used on platforms like Youtube, Facebook and Twitter. In this light, empowering communities seems like an alternative worth looking into.

There is active research happening in the domains of Machine Learning & Natural Language Processing, network modelling and Online Information literacy to tackle these issues. These are platform focused or individual focused approaches and are not centred around communities.

Whatsapp is in a place in its growth where it is now bigger than purely personal messaging and smaller than large scale group chat like Telegram and Discord. This comes in the form of a lack of channels(as compared to Slack/Discord), threaded replies(Slack) etc.

Whatsapp suffers from being a messaging application which is also being utilised as a content platform. It has issues pertaining to both its messaging aspect and its content management aspects. This project is a hypothetical design for Whatsapp. It is an adaptation of various moderation strategies that have been successfully adopted by various

online communities. The various constraints and challenges are outlined above and include the core tenants of whatsapp being a simple to use, privacy focused messaging application. The users for these tools are also much less tech savvy than users of similar tools on established platforms like Reddit, Wikipedia, Discord etc. The users are also mobile first and hence the tools need to be adapted as such.

Although followed in most other online communities; concentrating power to a select few in the group can potentially lead to tyranny and a feeling of powerlessness for others. This needs checks and balances to ensure that everyone is able to express their feelings and is able to mobilise their community to shift the status quo if needed. Again, this issue is harder to tackle for whatsapp because of a lack of central directory for available communities as present in Reddit, Mastodon Social etc.

1.6 Video Presentation

Some of the contents of this report have been covered in a concise video available at <https://www.youtube.com/watch?v=yeZ1ui6Gd5s>.

2. Secondary Research

The secondary research was done around two major themes. One was how content focussed platforms deal with the large user base and control the content vs how messaging applications manage ongoing conversations effectively.

2.1 Platforms: Content focussed

2.1.1 Reddit

Reddit pitches itself as the front page of the internet. Millions of people use it for everything from finding funny memes to participating in social activism. It has varied communities known as subreddits which are dedicated to a variety of topics. They are denoted by r/NameOfSubreddit. A few topics are news(r/news), cute animals(r/awww), Donald Trump(r/The_Donald), TV shows(r/BreakingBad, r/GOT), memes(r/dankmemes, r/memes), pseudo-profound statements(r/ShowerThoughts), personal disaster stories(r/TIFU, r/AmITheAsshole). Each subreddit has explicitly defined rules and members are expected to follow them while posting or commenting. Members submit posts containing links, images and videos they think might be of interest to the community. It has a voting(upvote/downvote) mechanism through which readers can help surface the best quality content organically. By itself, this system works well but is not enough at all. Members' posts and comments are

tightly regulated by a group of moderators (also referred to as mods). They are unpaid volunteers who put in this effort to help build and maintain communities that they care about.

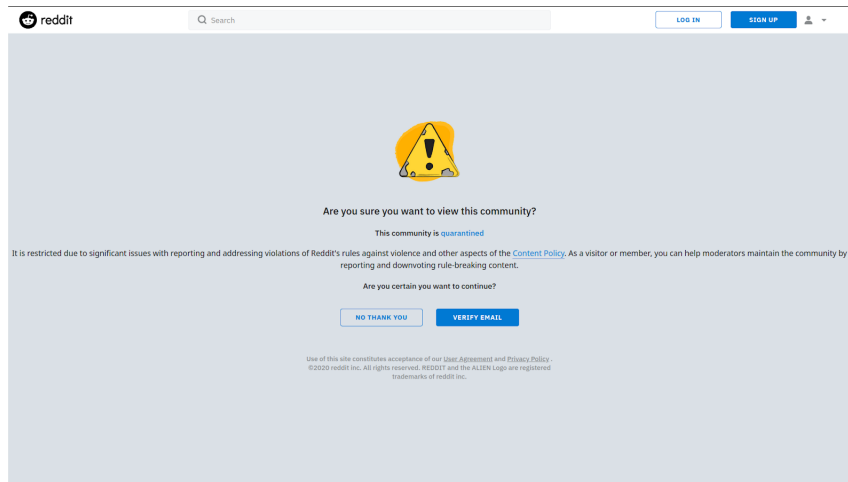
One recent story about what happened in r/worldpolitics points to the importance of the service mods offer on Reddit. r/WorldPolitics was a tightly regulated serious subreddit to begin with. In May 2020, after internal beef between the mods and the members, it suffered from a case of mods refusing to moderate and changing the subreddit policies to allow any posts. Within hours the subreddit was flooded with memes and NSFW(Not Safe for Work-Nudity, Porn etc.) images. To this date, the subreddit is in a similar state and little to no content related to World Politics actually shows up. Please visit [Reddit /r/worldpolitics Moderation Controversy(Know Your Meme, n.d.)] to read the full details.

There are various tools available to the mods to perform their job with ease(Reddit, n.d.):

- **Moderation bots:** Automated scripts that can do minor things like check broken links to doing preliminary filtering for hate speech
- **User Management:**
 - **Banning & muting:** Members can be banned from posting permanently or for a time period. Muted members are unable to send messages to admins for 72 hours
 - **Approved submitters:** Members that do not need to get each post verified.

- **Moderators and permissions:** Different levels of administrative powers can be given to different members of the community.
- **Flair - members & posts:** Subreddit specific tags that help other users distinguish between various types of members and posts within the community.
- **Post requirements:** These are requirements that posts need to adhere to. These may include things like having appropriate titles or only posting stories/pictures that you own and are original content(OC)
- **Removal reason:** While banning, a mod can also choose to provide a reason for the ban(Breaking rules constantly/ spamming/ karma farming etc).
- **Lock, OC, NSFW, Spoiler:** These are built-in tags that the reddit platform provides. They are used to categorise posts.

Apart from this, Reddit is very hands-off as a platform. They provide the necessary tools to administrators and moderators but do not prescribe their usage. Reddit has been at the centre of controversy regarding their stance. This was due to problematic subreddits like r/The_Donald(For breeding white supremacy) & r/TRP(for breeding misogyny) among others which reddit decided to quarantine rather than shut down even though they violate its content guidelines.



2.1.2 Facebook Groups

Are a feature within Facebook that help users connect to people they may/may not know personally as friends. Groups have a dedicated purpose/topic of interest or just a convenient way to share content with a specific audience.

Facebook offers various tools to help maintain a group(Facebook, n.d.):

- **Membership approval questionnaire:** Prospective members need to answer a bunch of questions for the admins to gauge their interest
- **Remove/block user:** A member can be removed permanently along with their posts or blocked temporarily
- **Admin vs moderator:** Admins have a higher privilege than moderators. They can alter basic group properties while moderators mainly moderate the content within
- **Approve each post:** Causes nothing to be posted without approval.

2.2 Messaging Applications: Communication focussed

2.2.1 Discord

Discord initially started as a medium for gamers to communicate while playing online. It has native support for Audio channels alongside the more commonly found text and media channels. It has since been reappropriated for other uses like coordinating between open source project contributors, Engaging with MOOC(Massive Open Online Course) students etc. It is designed for larger groups while also allowing smaller groups to function well. It does not offer end-to-end encryption.

Discord is geared towards users who are comfortable with technology. Inspired by IRC(Internet Relay Chat), it uses the metaphor of Servers and Channels to organise its communities. It also provides API(Application Programming Interface) access for use by developers/advanced users to build tools that fit their needs.

Discord has a powerful method to manage group communication:

- Allows the creation of **Roles with custom permissions**
 - These roles can then be assigned to various members of the server in different channels
 - This allows easier management of permissions for various categories of users
- **Permissions include:**



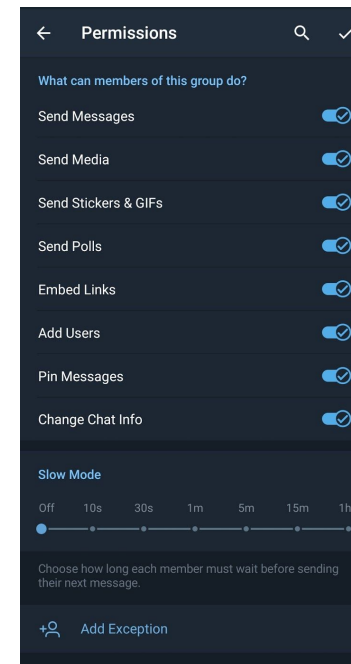
- **Slow-mode Cooldown:** Allows conversation to be slowed down by restricting each user(those without special permissions) to only a few messages per time interval.
- **Moderation bots:** It allows the use of bots(automated scripts) through the API to help manage the server. They can have various uses like reminding people to be polite, or removing certain types of content, preventing spam etc.

2.2.2 Telegram

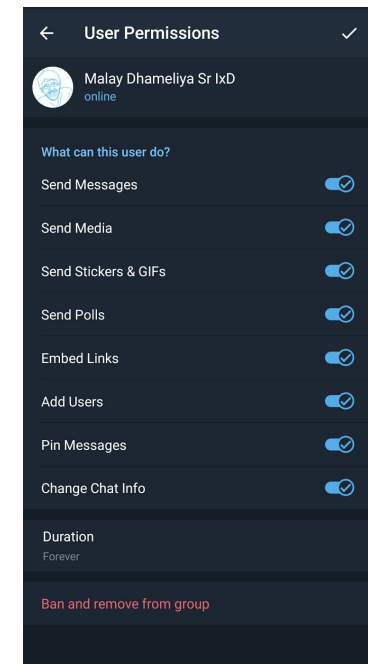
Telegram is an easy to use open-source IM application designed for the general public. It is generally used for large groups(>200 members). It does not offer end-to-end encryption by default. Channels in Telegram are used to refer to one-way communication broadcasts that users can subscribe to. These are used for various purposes by News organisations, Government bodies, celebrities, politicians etc.

It has various options for managing the groups:

- **Bots:** For various tasks not including moderation.
- **Permission management:** It offers a simple system of managing permissions for group as well as individual levels.
- **Slow Mode:** Limit the frequency of messages in the group



Group Level Permissions



Member Level Permissions

2.3 Insights from BBCs Duty, Identity, Credibility Report

(Chakrabarti, Stengel, & Solanki, 2018)(*Excerpts in Italics*)

This report was done by the BBC to throw light on Whatsapp usage in India. It is based on a mix of qualitative studies(Auto Ethnography & Semiotic Analysis) and quantitative studies(Big data/network analysis & News Scan Topic Modelling). It was the turning point of this project's transition into the current brief. Some excerpts from this report are being added here as is, to help convey the motivations and decision making factors for the project.

2.3.1 Content norms: Summary

- Groups operate under unsaid rules regarding content
- Sometimes, there are also explicitly said agreements
- Admins and individual members actively police content that they feel does not belong in that group, given the composition or the stated intent behind the setting up of that group

“I am from the Kanyakubj samaj, we have like minded people, with similar beliefs, so we share many things on that without thinking. I don't do it on Raipur Doctors group” (Male, 34, Raipur)



Image 5: Response to a posted message: “Please do not share such messages. Group admin please pay attention to this message

Image from pg 38 of report

2.3.2 Off Topic Conversations

“... You don’t want, but still, the posts are coming through, right? The people keep sharing it, it’s very irritating...Constantly, you will keep getting. There are groups in that people have their political affiliations. The regular ones have come, it’s Okay. Political things keep coming. It’s very irritating.” (Male, 38, Mumbai)

“Interviewer: Do you discuss only cricket related topics here?”

Respondent: Yes. It tells us about the trip series match, timings, the trophy picture is put up. I have a few friends in this group who are in favour of Modi and a few who are against him they constantly have fights over this. I just comment once or twice here. If it goes on then the Admin of the group tries to calm them saying this is a cricket group, no Kejriwal or Modi discussions. At times

their arguments heat up so much that till 1am the phone keeps buzzing. I have to mute the group. I have not made a group to discuss politics specially” (Male, 41, Delhi)

“Suppose I am not giving RSPC exam and somebody writes to me that this is very important please download it, I downloaded it but that was no use for me, so I am losing my MB, my memory is full, then I get some irritation.” (Female, 25, Udaipur)

The last quote above also points to another peculiarity in the Indian Whatsapp ecosystem, users prefer images or image heavy messages. Long text forwards and videos are less preferable due to a lack of storage and time/effort involved.

Due to the frustration of dealing with unwanted messages and breach of group decorum, admins & members are actively involved in the act of calling out offenders, but there is very little that admins or group members can do except kicking them out. Kicking someone out is seen as an extreme last resort measure.

3. User Stories

3.1 Cricket Group

Some cricket enthusiasts in a neighbourhood formed a group to discuss the latest developments in cricket. This involves everything from the political aspects of the BCCI's decisions to the performance of individual players. They also use it to organise matches on holidays. Some members also see it as an escape from the Covid-19 crisis. Others share health tips and infection stats in a bid to be informative. This sometimes descends into heated political fights about the government's working. The pandemic messages and off topic debates annoy bystanders as well as the admins. The admins are the younger players who created the group and do not approve of this uncle attitude. They are here for cricket and not politics. The interested members are free to form a new group to discuss politics.

As the admins, they want to keep the conversation relevant to cricket, so that all members can enjoy discussing in the group.

3.2 Class Group

There is a class group that serves multiple purposes. It acts as a central place to send academic updates, as a discussion forum for those updates, as a medium of talking about the latest movies/government policies and a way to coordinate plans to go out etc. While all discussions are welcome in the group, some are more pressing and time sensitive. For example the decision to ask for an extension on an assignment deadline. Discussions can get taken over by a vocal minority sending a barrage of messages. If there are opposing points of view voicing their opinion, the discussion gets chaotic. Instead of bothering to read through hundreds of messages, other members either repeat the similar statements, or ask questions that have been answered previously. The accelerated nature of conversation also makes people put hastily written responses rather than making an articulated point. This derailing is of concern to multiple people, who have decided not to engage in any such discussions at all.

The group members want to have civil discussions so that decisions can be taken quickly with the active involvement of more people.

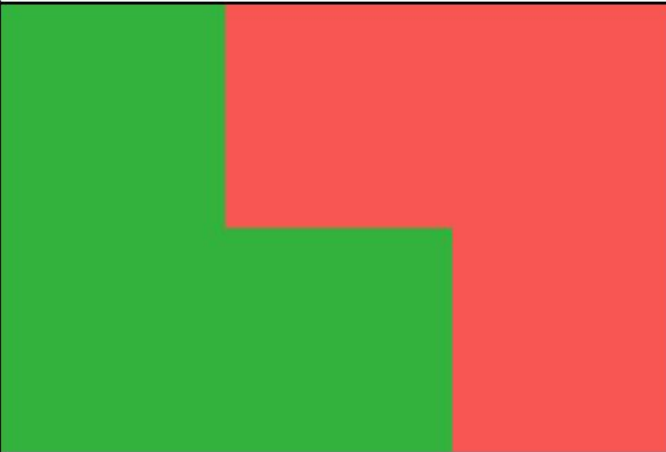
3.3 Online Learning Group

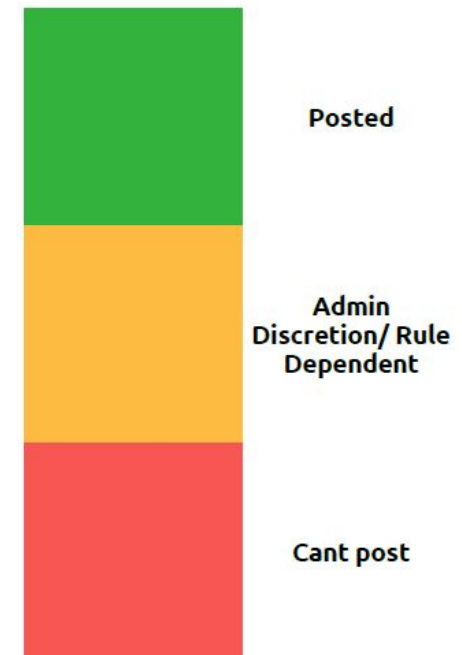
A teacher has recently moved her class online due to the closure of schools. She teaches Mathematics to 12th graders in a private school. She and her students have used Whatsapp for their personal communication needs for a few years and are comfortable with the core functionality. Like most classes, there are students who want to study, there are those who don't and others who couldn't care less either way. She sends out zoom links for her classes. To keep the students engaged, from time to time she also sends out fun whatsapp quizzes to spark discussions within the group. These have mixed results. Some students send interesting Youtube links related to the topic while one or two end up spamming the group with memes and forwards.

As a Teacher she wants to encourage healthy posting while discouraging/stopping spamming, so that she can foster a better learning environment in her whatsapp group.

4. Proposed Changes

Diagram-1: Current state system of Whatsapp

Moderation Mode (While Posting)	Admin Messages only			
	Allow Everything			
Participant Label (Pre Posting)		Admin	Normal	Removed



Admin Power (Post Posting)		Delete Member Messages					
Moderation Mode (While Posting)	Admin Messages only						
	Moderate each message						
	Moderate flagged messages						
	Allow Everything						
Participant Label (Pre Posting)		Admin	Sub-Admin (Elevated Posting Rights)	Normal	Flagged (Restricted Posting Rights)	Read-only	Removed

Diagram-2: Proposed state system of Whatsapp

4.1 Current v/s Proposed states of Whatsapp

Before getting into the specifics of what the differences between the two state management systems is, we'll talk about how they are analysed. The diagrams are vertically broken down as the three different stages of posting a message that succeed each other. The bottom one representing the permissions that different participants may have. The middle layer represents the different kinds of content flow moderation modes available to the administrators of the group to choose from. The coloured boxes represent how these modes interact with the participant permissions. The top section (only present in Diagram-2) represents the possibilities that exist after a message has been received in the group.

4.1.1 Current State

(Please Refer to Diagram-1) With regard to group posting rights management, Whatsapp offers very little. As illustrated through the Y-axis of the matrix, a group participant can move only between being:

- **Admin**, whose permissions include:
 - Adding/removing participants
 - Controlling who the other admins are
 - Be the only one with ability to post in certain situations
- **Normal** (Unlabeled), whose permissions include:
 - Posting when everyone is allowed to post
- **Removed** (Cease to be a participant)

On the X-axis, we see that there are two modes available to control the flow of messages.

- **Allow Everything**: The default and places no restriction on the flow of content.
- **Admin Messages only**: Allows only admins to be able to post anything in the group

Another feature that is offered in this regard is the ability to delete only your own messages, we'll see later how this offers very little in terms of being useful for group management.

Shortcomings of current features:

The current methods on offer are limited in their ability to be used effectively as tools for group management. They are crude and offer no fine grained control over permission access or with regards to content. Because of their hammer or nothing approach, it results in very restrained use of these tools, if at all to manage participant posting behaviour. This lack of usefulness is detrimental to the larger goal of maintaining group norms where most people feel comfortable. At the very least, even the ability to delete another participant's message does not exist, no matter how un-welcome it might be in the group. For those worried about the free speech implications of it, do not worry. We will be discussing that later.

4.1.2 Proposed State

(Please refer to Diagram-2) In this state we propose to increase the granularity that Whatsapp offers in terms of participant permission management and introduce rudimentary content moderation as well. Content moderation(Referred to as Review in implementation) is achieved through the introduction of the concept of flagging(explained later). Starting from the bottom section again, we first focus on the different levels of participant permissions are:

- **Admin**, whose permissions include:
 - Reading all messages
 - Adding/removing participants
 - Managing participant permissions
 - Managing group content rules
 - Accepting/rejecting messages
 - Deleting other participant's messages
 - Being the final authority on content in most cases
- **Sub-Admin**, whose permissions include:
 - Read approved messages
 - Elevated posting rights-They can post any messages regardless of the moderation mode, except for 'Admin messages only' mode
 - Post a link to a new group regardless of moderation mode
 - Trigger Slow/Conversation mode

- Reporting the Admin/Group
- **Normal** (Unlabeled), whose permissions include:
 - Reading approved messages
 - Posting with restrictions depending on moderation mode
 - Reporting the Admin/Group
- **Flagged** (Restricted Posting rights), whose permissions include:
 - Reading approved messages
 - Posting only through admin approval
 - Reporting Admin/Group
- **Read-only**, whose permissions include:
 - Reading approved messages
 - Not posting at all
 - Reporting Admin/Group
- **Removed** (Cease to be a participant)

Flagging

Flagging is introduced to create an easier mental model for content moderation. Through the use of different filters, messages that are potentially disruptive for the group norm can be flagged or banned automatically. The decision to Allow/Flag/Ban messages through a certain filter depends on the group settings that the admins choose. The group settings are visible to all group participants. This is a

framework that can be expanded in the future to cover complex topics like Fake News/Hate Speech etc using advanced techniques like on-device ML and NLP techniques like Participant Profiling, Emotion analysis and Topic Modelling.

Currently, we are proposing three filters which can work seamlessly with existing privacy protections in Whatsapp.

- **Flagged User:** All the messages from such users are flagged for review
- **Forwarded/Super Forwarded messages:** This allows control over the possibility of sending forwarded/super forwarded messages into the group. This relies on message metadata that Whatsapp currently possesses.
- **Blacklisting:** This contains a number of wordlists selected by the admins. These editable word lists contain multilingual-words related to a topic. For example, a Coronavirus Wordlist will contain words like Covid, Coronavirus, Covid-19, Wuhan Virus, SARS-Covid-2, कोरोना वायरस, कोविड-१९, कोविड-19 etc. Messages are simply checked against these words to find if there are matches. In case a message is found with matching words, it can be flagged/banned. It is a simple to understand process that will help users understand, adopt and tweak it more readily. On the other hand, it will have both false positives and false negatives and hence is mainly suitable only as a starting step in this direction.

Other Features

Two additional features are being proposed here-

- **Slow mode/Conversation mode:**
In case of a low level of moderation in the group, the slow mode works by limiting the frequency of messages a participant can send. Can be used in cases when a discussion is going extremely fast.
In case of a high level of moderation in the group, the admins see the option of allowing conversation to flow freely for a fixed time period. This works by allowing all the participants to post without needing approval. This can be used to hold discussions/invite comments in an otherwise restricted environment.
- **Media Restrictions:** Choose what media types are allowed to be sent to the group. This is based on the comments from the BBC report and would prove useful in situations involving users with limited storage.

4.1.3 Free Speech Implications

There is a tension between absolute free speech and reasonable restrictions on speech that exists both in our society and online interactions. Here we will try to unpack some of that argument with respect to these tools and suggest a features that help counteract admins misusing their discretion.

The restrictions on speech here are akin to what a person might feel in polite company before saying something untoward. This is the tradeoff that we implicitly agree to during group formation. These norms evolve over time and are dynamic. This tool only attempts to make these group norms more explicit, because the unembodied nature of online communication makes it much harder to negotiate implicit norms. These restrictions do not apply to one-on-one communication. You might have something to say, but a group is not bound to hear it if they feel it's not welcome.

The proposed tools are just an option that admins can choose to exercise rather than a default. Having said that, the potential for abuse is always there. The hope is that group members can coordinate amongst each other to set and maintain the group norm using both the group as a forum and personal messaging. It is not always that this will be possible in all settings or for all people. These are the suggested features to deal with this -

- **Invite to new whatsapp group:** All participants will be able to send invitations for a new group. The ability to do this is a step in balancing out some of the admin's power to oppress.
- **Improved Reporting feature:** The current report group feature in Whatsapp is not very useful as it takes you out of the group and reports the group to whatsapp, which due to end-to-end encryption is unable to do much about it. This is now broken down into two. One, to report any abuse of power/bullying by the admins to the rest of the group. This will help call out unpleasant behaviour explicitly. This is restricted to being limited to one report a day per participant to prevent harassment through excessive use. The second, to report the entire group in case it is filled with objectionable content. We have not designed for such a case, but propose a few characteristics. The function should generate a verifiable, uneditable & shareable document/archive of the group's complete messages and media. This can be achieved by the app signing the document digitally using a part of the group's encryption key. Its uneditable and complete nature would serve to prevent misappropriation of messages. This would be a step up from the prevalent method of screenshotting that was also used in the recent "Boys Locker room" incident(Ramesh, 2020).

4.1.4 Non-Proposal

Bots: They are not included in the proposal for multiple reasons. One of the primary reasons is privacy. Currently, the messages are only readable on device. While some bots can operate locally, others might relay the messages back and forth in communicating with a server. Allowing bot access is a serious threat to the core feature of whatsapp being Privacy centric. Another concern is that while bots are used on other platforms and messaging applications, they tend to have a much more tech savvy user group. Whatsapp is designed for the general public and the introduction of bots might be a hurdle for some.

AI/Third person based moderation: Large platforms like FB and Youtube have advanced ML models that help keep content to their community standards but it is a very data intensive practice. This invasion into a personal messaging application is hard to justify. Human moderation regularly costs companies large sums of money and results in poor working conditions for the people who have to deal with the mental trauma of sorting through objectionable content for extended periods.

Voting/polling System of any kind: Any kind of voting system has been avoided because of the passive nature of messages in chats. This also helps to stay true to its simple messaging application premise.

Channels: While channels are supported by advanced communication tools like Slack and Discord, introducing channels will add a layer of abstraction within groups. This is detrimental to the simple model that whatsapp operates with.

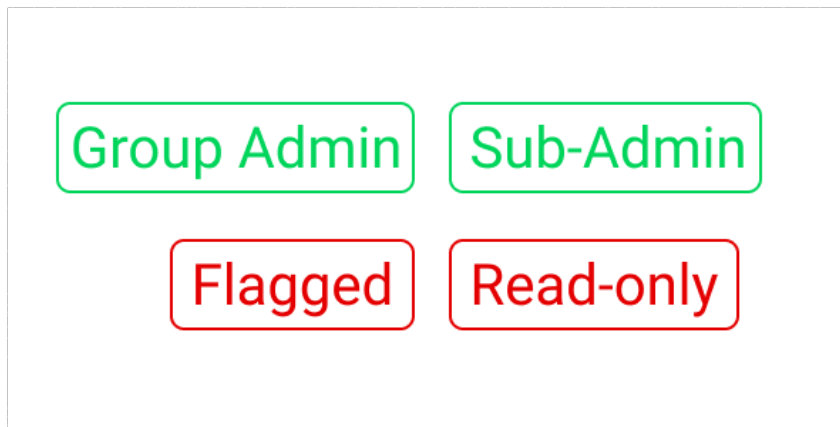
Customizable roles: This places additional responsibility on users to come up with roles that might be suitable for their groups. Instead, the decision was made to stick with predefined roles('Admin', 'sub-admin', unlabeled, 'Flagged' and 'Read-only') which are in a descending order of privileges. This makes it easier to move people up and down this ladder based on their actions in the group.

4.2 Implementation

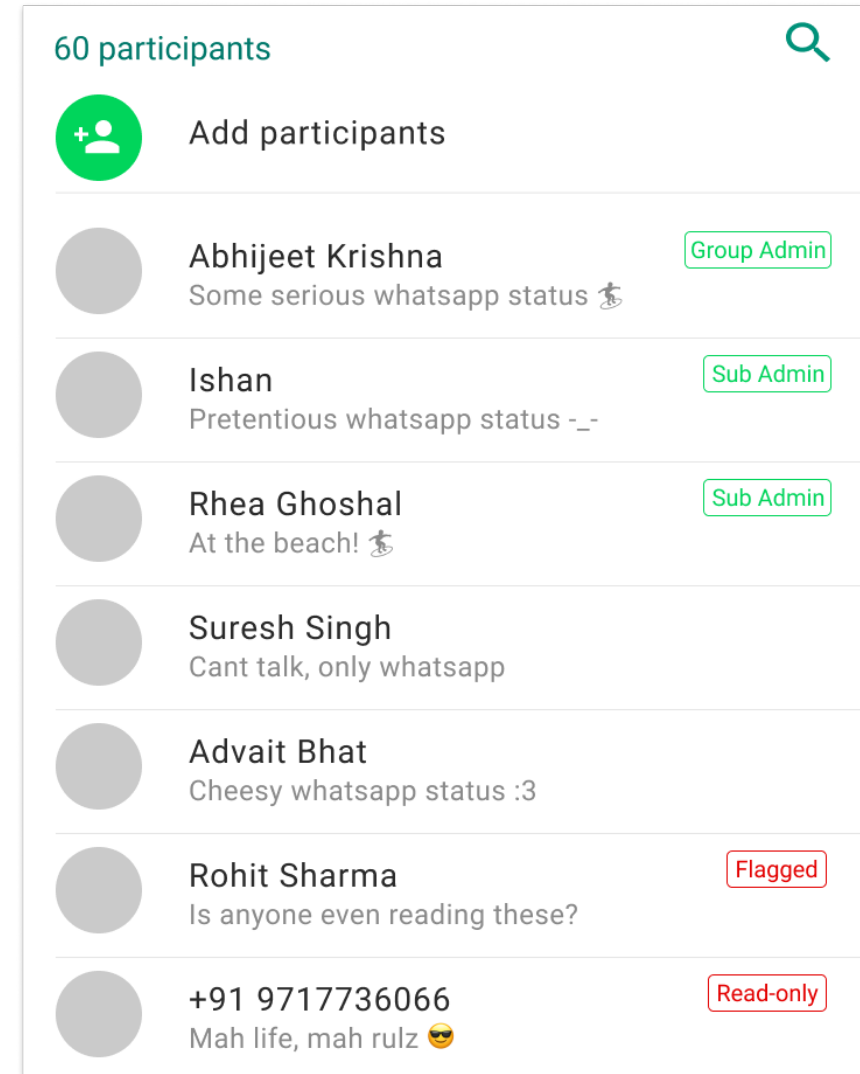
The visual design for additional components is based on the existing Whatsapp Design language. This includes Typography, Colours, Iconography and Layouts.

Note: Please visit <https://www.youtube.com/watch?v=yeZ1ui6Gd5s> and skip to 8:42 for a video walkthrough of the details displayed here.

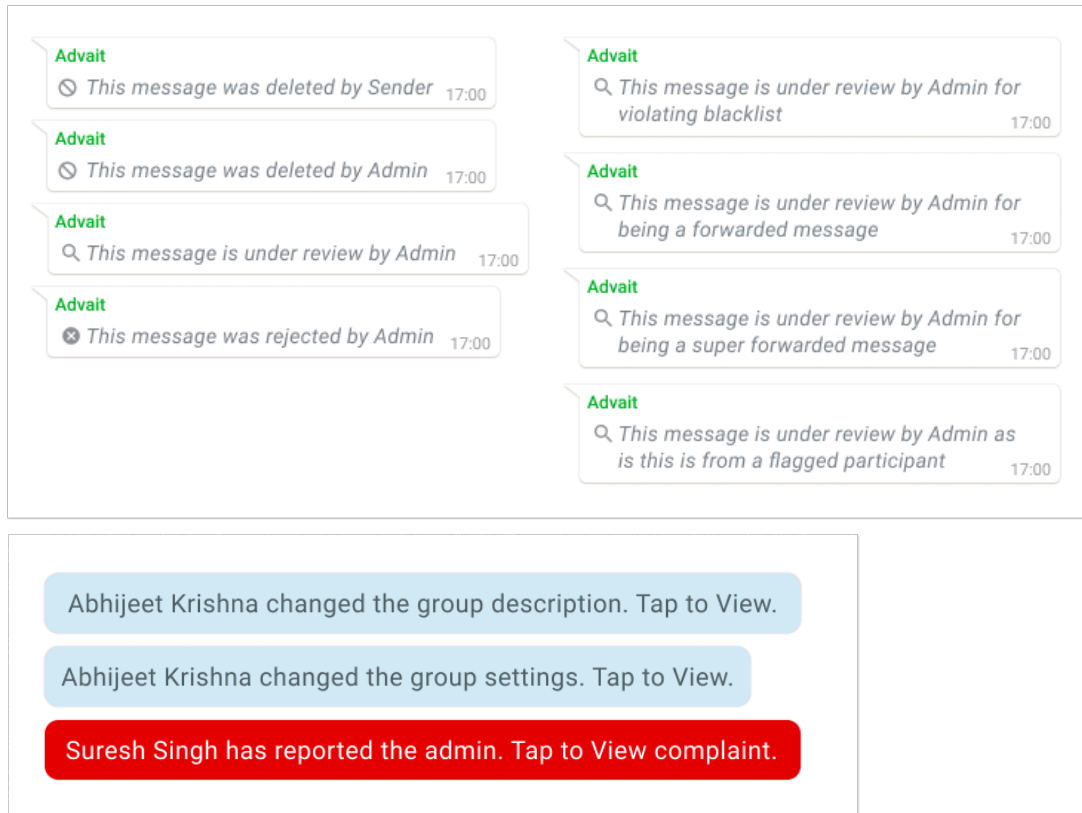
4.2.1 Participant Tags



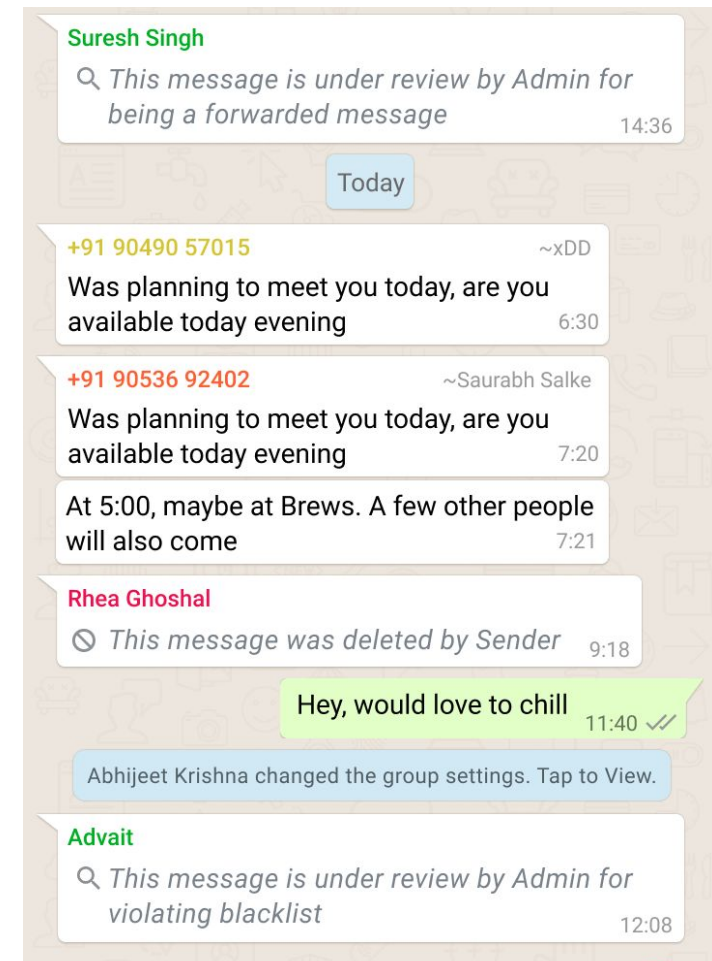
The Participant tags are based on the Group Admin tag that is available in the original Whatsapp.



4.2.2 Message Status Prompts, Group Status Prompts



The information about the current status of the message in the review cycle is shown by the message status prompts. They are based on the design of the “This message was deleted” prompt that Whatsapp already has. The prompt indicating the message to be under review comes with multiple alternatives depending on the reason for the flagging. The participants are notified of any change to the group setting/reports against the admins through the Group Status



prompts. This is based on the “changed group description” design. Both categories of prompts are designed to introduce transparency into the process of moderation.

4.2.3 Slow-Mode/Conversation Mode

Slow Mode

Choose how long each participant must wait before sending their next message



Conversation Mode

Choose how long each participant can send messages without needing approval



Due to a lack of any existing sliders in the existing Whatsapp UI, a slider was adapted from Google Material Design to fit the Whatsapp visual design. The cards are taken from the card style in the group details pane.

4.2.4 Group Settings

(Images on following page) Group setting is the menu where most of the newly introduced features are controlled from. Before this, it was used to set group info editing and message sending permissions. Now it is utilised to set moderation mode, Filter properties media restrictions and Admin delete powers. The filters will be inactive (greyed out) if the moderation mode is set to no review. This same page is used by non-admins to view the current rules.

The visual design is based on the button and overlay popup pattern found in the original.

4.2.5 Blacklist

As the blacklist feature is a more complex concept, it required additional steps to offer full functionality. The mockups for which are on the following page. The wordlists are editable to add or remove words as per the regional differences in use. In the case of the blacklist, non-admins can see the topics but not the wordlist itself to help prevent gaming the system.

The visual design is based on the text-box in a popup found in the Feedback section of Whatsapp.

Edit Group info

All participants

Send Messages

All except Read-only

Messages requiring review

Flagged

Forwarded messages

Allow

Super forwarded messages

Ban

Media Restrictions

None

Blacklist

Active

Can Admin Delete Messages

Send messages

Choose who can send messages to this group

☐ All participants

☒ All except Read-only

☐ Admin & Sub-Admin

☐ Admins only

CANCEL OK

Messages requiring Review

Choose which messages will need to be approved by admin before posting to group

☐ None

☒ Flagged

☐ All

CANCEL OK

Forwarded messages

Choose if forwarded messages can be sent to the group

☒ Allow

☐ Flag

☐ Ban

CANCEL OK

Super Forwarded messages

Choose if super forwarded messages can be sent to the group

☐ Allow

☐ Flag

☒ Ban

CANCEL OK

Media Restrictions

Choose which media is allowed in the group

☒ Photos

☒ Audio

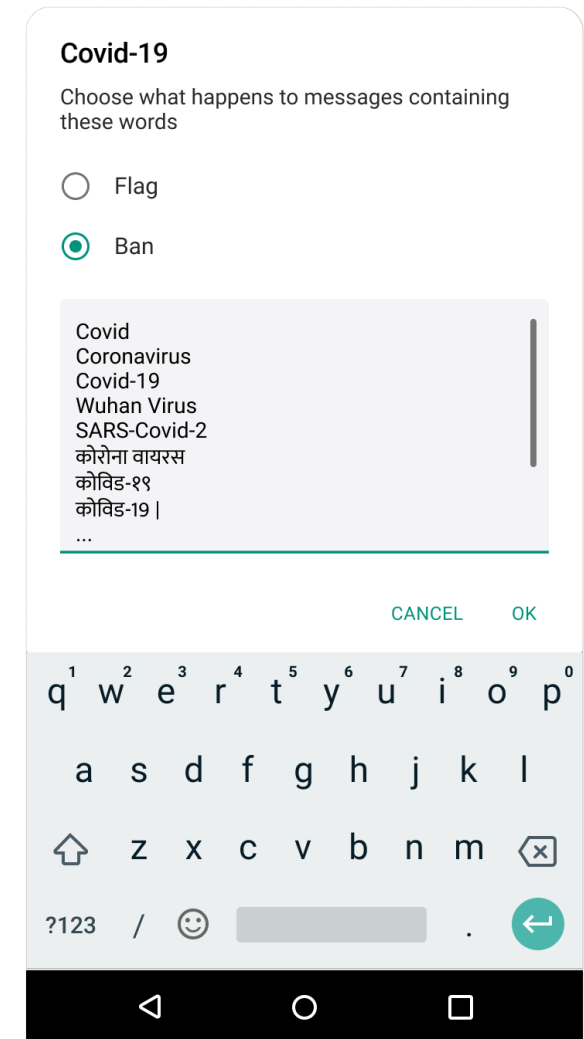
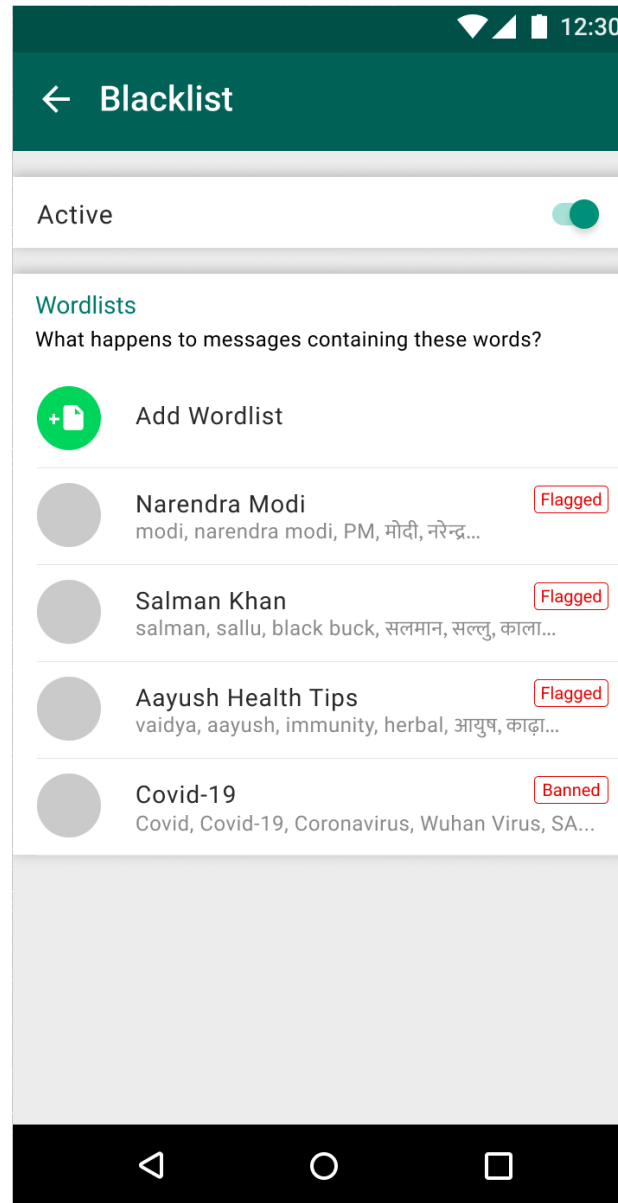
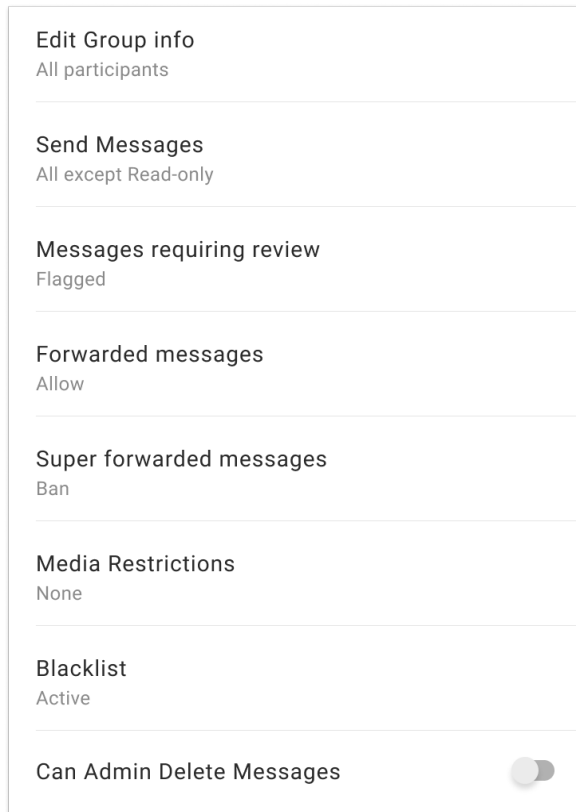
☒ Links

☒ Videos

☒ Documents


CANCEL OK

Overlays



4.2.6 Report

The report feature has been redesigned to meet the needs for accountability better.


Report

Report

Make a complaint against Admins in the group. Or report the group to Whatsapp

You can only report once per day

☒ Only Admins

☐ Group

☒ Report Anonymously

CANCEL
NEXT

Report Admin

Please describe why you want to report the admins, be as specific as possible.

Multiple times the admins refuse to post my message. They say I have nothing useful to say and have banned me from posting anything at all. This was done without reading through my messages. This makes me extremely uncomfortable. This is unfair and I feel bullied. Other members and sub-admins are requested to please look into it and demand the admins to stop this or provide reasonable justification. |

CANCEL
OK

¹q²w³e⁴r⁵t⁶y⁷u⁸i⁹o⁰p

a s d f g h j k l

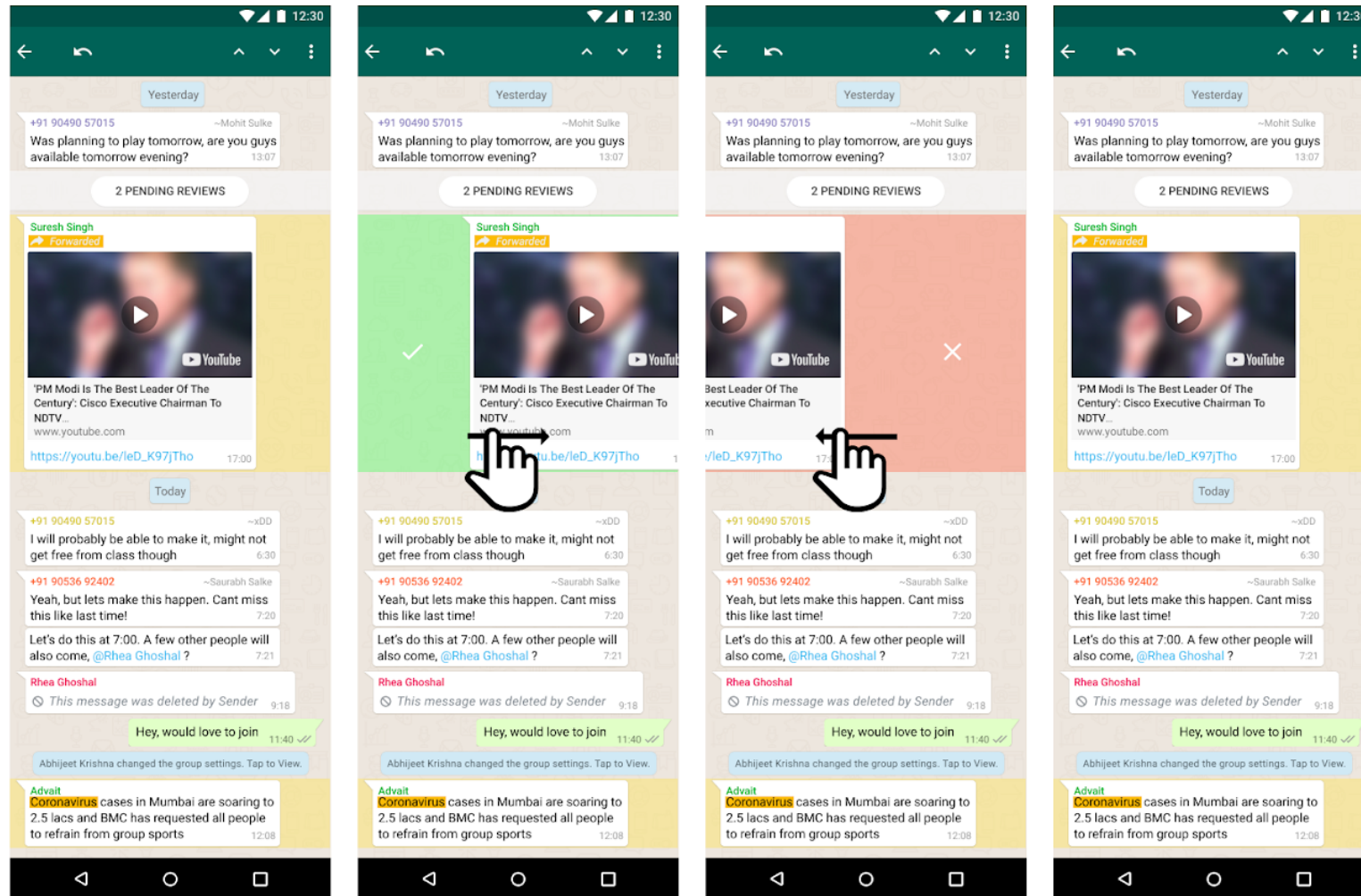
↑ z x c v b n m ↵

?123 / 😊 . ↩

4.2.7 Review (Moderation)

Moderation is called Review in the implementation so that users can hopefully understand the functionality more easily. The review section is chosen to be in the chat format to keep track of the conversation-context in which those messages were sent. This is in stark contrast to approaches used by content focussed platforms, since they deal with independent posts and not interlinked conversation. The interaction for approving and rejecting messages is chosen as swiping right and left respectively. This interaction is already used in Whatsapp to reply to messages. It also features an undo option. The interaction chosen for going to the next message needing review is the same as search result browsing. On long pressing the message under consideration, other than Accept/Reject there are options to Reply privately to the sender, to Flag/unflag sender and change the sender tag to read-only. This is done to make a variety of options available in one place replacing the need to browse deep nested menus or navigate to other menu places.

Future Changes: This can be thought of as a starting point for the Review mode. Currently, it is designed while keeping usability in mind. In the future it can be modified to fit other issues as well. For example, it can be designed to be deliberately slow to encourage reflection when dealing with issues like fake news.



Approve Message

Reject Message

Reply Privately

Unflag Suresh Singh

Make Suresh Singh Read-only

Conclusion

Whatsapp has proven itself to be a double edged sword. It offers easy communication ability that has contributed to its widespread adoption and use as a content platform. This simplistic handling of affordances has led to a lack of agreement on group norms within participants. This is a source of annoyance and clutter in the groups.

To address these issues, the project borrows the approach used by content platforms like Reddit and FB Groups. It empowers group admins to regulate the content in the group. To aid that process, the concept of filters and flagging is introduced. The challenge was to work within Whatsapp's constraints of encrypted messages, simple to use application, mobile first usage and a user group with limited tech skills.

The project throws light on one of the ways these issues can be mitigated. While best practices and existing design patterns have been reappropriated, this is speculative at the moment. The nature of the idea itself does not lend itself to testing very easily. These features have a clear social component that works well only in context. To be able to test it out of whatsapp's context (by making a similar app) while valuable will not offer the insights needed to establish their effectiveness. This is near impossible to achieve in the current Whatsapp app through Wizard of Oz methods since Whatsapp does not provide API endpoints or modifications of any kind to the app.

One of the most effective ways to test this would be, if Whatsapp would allow or make changes to the code and allow the modified app to run on the test participant's in regular use for some time period. This is under consideration and a proposal is being drafted to be sent to Whatsapp regarding the testing and integration of this project.

This is still the first step in the broader battle of tackling fake news and hate speech on closed platforms like Whatsapp. This is not even close to the required levels of intervention, it addresses a more basic need instead, that of the digital infrastructure through which interventions can be delivered in the future. This will act as an entry point for future researchers and designers to insert filters and models that take various factors into account and enable the weeding out of such content. They can be further expanded to tackle bad faith admins. This would again give rise to a tension between regulation and censorship; which will have to be addressed by the designers and policymakers at that juncture.

Reflection

This project was one of the hardest I have taken on for multiple reasons. Considering there are two parts of a project, the domain and the medium, I chose a medium I was familiar with i.e. game design and an unknown domain i.e. fake news. I had very little knowledge of it except for feeling that it is a pressing problem that needs solving. Years of listening to Professors explain what constitutes a design problem perhaps finally got driven home only after this experience. In my defence, I still feel one understands the nature of a domain only after spending sufficient time in it (especially a domain this loosely known) and calling it a day before that is not doing justice to it either. Another thing I learnt about myself was that I was unable to pivot out of this topic early on because I both believed something could be done and also had a fear of failure. On the surface, I was okay with coming up with any end product without regard to its effectiveness, but the fear was much more internalised. It stopped me from coming up with 'anything' or 'going wild' with my imagination. If I was not completely comfortable with trying something radical and failing miserably, did I not miss out on a core design-principle of embracing failure? It feels like these principles have been taught but not practiced. Failure has met with harsh criticism that was rarely constructive. This drive to have an end product served to discourage rather than provide the motivation to reach a tangible goal.

Being an undergrad in a design school, with a loosely understood topic, the argument should probably have been, no one knows better anyways, maybe this will work out. I couldn't get myself to do it.

I was attached to the idea that I could make things work purely because I was enthusiastic about them. With time, I opted for this much safer, much better understood topic to close this chapter of Bdes with. This is growth in a way, and a reminder that romantic ideas rarely serve well in the pragmatic world.

Bibliography

Chakrabarti, S., Stengel, L., & Solanki, S. (2018). DUTY, IDENTITY, CREDIBILITY: 'Fake news' and the ordinary citizen in India. BBC.

Facebook. (n.d.). From Facebook Web Site:

https://www.facebook.com/help/1686671141596230/?helpref=hc_fnav

FE, O. (2020, April 7). The Indian Express Group. From Financial Express Website:

<https://www.financialexpress.com/lifestyle/will-govt-punish-whatsapp-group-admins-members-for-sharing-coronavirus-jokes-check-fact/1921075/>

Know Your Meme. (n.d.). From Know Your Meme:

<https://knowyourmeme.com/memes/events/reddit-rworldpolitics-moderation-controversy>

Ramesh, M. (2020, May 05). The Quint. From The Quint Website:

<https://www.thequint.com/news/india/recovered-only-15-operational-cost-of-shramik-trains-railways>

Reddit. (n.d.). Reddit. From Reddit mods:

<https://mods.reddithelp.com/hc/en-us/sections/360000208432-Moderation-Tools-on-New-Reddit>

Roy, D. D. (2017, April 23). NDTV. From NDTV Web Site:

<https://www.ndtv.com/india-news/whatsapp-facebook-group-admins-can-go-to-jail-for-offensive-posts-1684001>

Sikdar, S. (2016, March 29). The Hindu. From The Hindu Web Site:

<https://www.thehindu.com/news/national/other-states/if-you-are-a-whatsapp-group-admin-better-be-careful/article7531350.ece>

Singh, M. (2020, April 27). Tech Crunch. From Tech Crunch Website:

<https://techcrunch.com/2020/04/27/whatsapps-new-limit-cuts-virality-of-highly-forwarded-messages-by-70/>

Whatsapp. (n.d.). About Whatsapp. From Whatsapp:

<https://www.whatsapp.com/about/>

YS, T. (2017, April 21). Your Story. From Your Story Web Site:

<https://yourstory.com/2017/04/whatsapp-admins-legally->

responsible-offensive-content?utm_pageloadtype=scroll