A First Look at COVID-19 Messages on WhatsApp in Pakistan

R. Tallal Javed*, Mirza Elaaf Shuja*, Muhammad Usama*, Junaid Qadir*, Waleed Iqbal†, Gareth Tyson†, Ignacio Castro†,and Kiran Garimella‡

*Information Technology University, Punjab, Pakistan.
†Queen Mary University of London
‡MIT
Email: *(tallal.javed, msds18051, muhammad.usama, junaid.qadir)@itu.edu.pk, †(w.iqbal, g.tyson, i.castro)@qmul.ac.uk, ‡garimell@mit.edu

***Abstract*—The worldwide spread of COVID-19 has prompted extensive online discussions, creating an 'infodemic' on social media platforms such as WhatsApp and Twitter. However, the information shared on these platforms is prone to be unreliable and/or misleading. In this paper, we present the first analysis of COVID-19 discourse on public WhatsApp groups from Pakistan. Building on a large scale annotation of thousands of messages containing text and images, we identify the main categories of discussion. We focus on COVID-19 messages and understand the different types of images/text messages being propagated. By exploring user behavior related to COVID messages, we inspect how misinformation is spread. Finally, by quantifying the flow of information across WhatsApp and Twitter, we show how information spreads across platforms and how WhatsApp acts as a source for much of the information shared on Twitter.**

***Keywords*—COVID-19, Misinformation, WhatsApp, Twitter**

I. Introduction

Social media apps like Facebook, WhatsApp, and Twitter have changed the way we communicate. The information disseminated through these apps has influenced our social and cultural norms in an unprecedented way. WhatsApp is one of the most frequently used and rapidly growing social media apps in the world with more than 1.5 billion users. WhatsApp is also ranked #1 for the average number of active users in the world per month [1].

WhatsApp has, therefore, become important for understanding social behavior and opinion formation. In many cases, the scope of the information shared in WhatsApp groups, is limited to a community or a country [2]. Recent studies [3] have shown that WhatsApp (like other social media platforms) is also used for the dissemination of misinformation. We argue that it is important to understand how this false information, either spread knowingly ("disinformation") or naively ("misinformation") influences opinion formation in different societies. This is particularly important in the global south, where despite users having low digital literacy, WhatsApp is the de facto mode through which users obtain and share information [3].

With this in mind, we perform a comprehensive analysis of COVID messages being propagated through WhatsApp in Pakistan. Pakistan is a major developing country, with approximately 37 million active social media users[1]. Building on the idea of how political parties around the world [4] are using public WhatsApp groups to reach their audience, we start by monitoring a large sample of public WhatsApp groups related to politics in Pakistan. Meanwhile, a major world event occurred: On March 11, 2020, the World Health Organization (WHO) declared COVID-19 a pandemic [5]. Our data also holds unique value, as it encompasses non-covid groups, and gives us insight into how COVID related content, organically gets propagated across public WhatsApp groups. Specifically, we explore the following research questions:

RQ1: What kinds of messages, about the pandemic, are being shared, on the publicly accessible WhatsApp groups of Pakistan?
RQ2: Is there misinformation related to COVID-19, and if so, to what extent and of which type?
RQ3: What is the general user behavior, and can we detect disinformation from it?
RQ4: What is the interplay between misinformation related to COVID-19 shared on WhatsApp and Twitter?

To explore these questions, we have collected data from 227 public WhatsApp groups starting January 10, 2020. To the best of our knowledge, this is the first dataset and analysis of COVID-19 related conversations from a country in the global south, involving multiple modalities (text and images) and multiple platforms (WhatsApp and Twitter). We begin our investigation by analyzing the content shared in the WhatsApp groups and filtering out the COVID-19 related content. The filtered content is then further divided into text, images, videos,[2] and other related categories. Using this data, we make the following contributions:

- We offer the first WhatsApp dataset consisting of discussions related to COVID-19 from Pakistan. The dataset includes texts, images and videos originating from 227 groups. The (anonymized) dataset will be made publicly available to the community.
- We show using extensive manual annotation that around 14% of the messages related to COVID-19 had misinformation about the pandemic.
- We perform a temporal analysis of misinformation related to COVID-19 propagation, across WhatsApp and Twitter, exploring how content is copied across.

[1] https://datareportal.com/reports/digital-2020-pakistan

[2] In this study, we focus on text and images, leaving video analysis for future work.

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

II. Related Work

A. (Mis)Information on WhatsApp

WhatsApp has been a source of major political misinformation and propaganda campaigns [6], [7]. Political parties have invested heavily in social media strategies by creating WhatsApp groups to reach WhatsApp users [8]. Surveys done in India and Brazil show that at least one in six users are part of one such public political WhatsApp group [9], [4].

Garimella et al. [10] provide tools to collect and analyze public WhatsApp group data at scale. Making use of these tools, various studies have shown the extent of misinformation and manipulation on WhatsApp [11], [12], [13], [14]. Particularly, Resende et al. [12] analyze doctored images to fuel smear campaigns against political rivals and the dissemination of misinformation through WhatsApp groups in Brazil. Garimella et al. [14] provide an analysis of image-based misinformation spread during the 2019 Indian elections and show that over 13% of the images contained misinformation. Melo et al. [15] provide a system for gathering analyzing, and visualizing WhatsApp public group data for identification of misinformation propagated in three countries: India, Brazil and Indonesia. Maros et al. [16] analyze audio messages shared on WhatsApp and characterize their propagation dynamics. The analysis is performed on 20K audio messages from 330 WhatsApp public groups and the results suggest that the audio messages with misinformation spread further more than the benign or unchecked audio messages.

B. Health (Mis)Information

A major focus of this paper is understanding the spread of health misinformation related to COVID-19. WhatsApp has been a major source of health misinformation especially during the pandemic [17]. This misinformation ranges from highlighting wrong symptoms to ineffective treatments. Jin et al. [18] reported a massive wave of misinformation on social media, especially on Twitter during the Ebola pandemic in Africa. More comprehensive details on how fake news about Ebola on social media applications is explored in [19].

With the ongoing surge in the COVID-19 pandemic a wealth of misinformation has already been documented. Sharma et al. [20] provide a dashboard for analyzing misinformation about COVID-19 on Twitter. They analyze 25 million tweets and provide a country-wise sentiment analysis of how people are reacting to COVID-19. Singh et al. [21] analyze Twitter-based misinformation about COVID-19 and provide insights on how the propagation of misinformation on social media is connected to the rise in the number of COVID-19 positive cases. Kouzy et al. [22] analyze Twitter-based misinformation about COVID-19 and report that tweets having the keyword "COVID-19" contains less misinformation and tweets with keywords "2019-ncov" and "Corona". Cinelli et al. [23] provide a comprehensive analysis of the use of different social media platforms in the COVID-19 pandemic. They analyze Twitter, Instagram, YouTube, Reddit and Gab, providing a review of how the discourse on these applications is evolving. They also explore the propagation of misinformation from different questionable sources in social media.

C. Our Work's Novelty

To the best of our knowledge, there does not exist any work analyzing COVID[3] related discussions on WhatsApp. Since WhatsApp is arguably the most frequently used application in the world, it is important to study it to see how people are using the platform during the pandemic and how the platform facilitates the spread of COVID-19 misinformation. Although prior work has focused on misinformation spread via WhatsApp in Brazil and India, we are the first one to study misinformation on WhatsApp during a major pandemic. Furthermore, our analysis is focused on Pakistan, which has a thriving Muslim religious identity, which allows us to see how religion plays a role in the context of public health. In contrast to the majority of prior work on misinformation, which focuses on textual analysis, we also provide a detailed analysis of images related to COVID-19 and study the information spread across WhatsApp and Twitter both for text messages as well as images.

III. Methodology

In this section, we delineate our data collection & annotation methodology, and discuss the related ethical issues.

A. Data Collection

WhatsApp allows its users to create public and private groups. The public groups can be joined by any user of the platform, typically through an invite URL of the form chat.whatsapp.com/*. These URLs are frequently shared via other social web platforms (e.g., Facebook, Twitter) to invite third parties to join.

Selection of groups. To compile a list of relevant public groups, we looked for chat.whatsapp.com links on Facebook and Google to find group invite URLs. We specifically targeted the popular political parties of Pakistan as these groups tend to be more active and give an idea of the political sphere. Hence, "WhatsApp" along with political parties' names and slogans were used to search for public groups.

Based on the above parameters, we compiled a list of 282 public WhatsApp groups. In order to ensure the quality of groups, we manually discarded groups that were unrelated. For instance, if a group's profile picture, group name or bio did not contain any relevant information (political aims/motivations) then it was removed. We further removed groups which were buying and selling things, and did not have any organic interactions/messages. This left us with 227 public groups, on which the analysis was done.

In order to find these groups, a set of queries, search engines and filters were used. These queries can be found at https://cutt.ly/8yXhxBd. We also plan to release our anonymised dataset once the paper is accepted to encourage further research on WhatsApp data from Pakistan.

WhatsApp data collection. To join and get data from the groups, we used tools provided by Garimella et al. [10], which uses the Selenium Web Driver to automate the joining of the groups. WhatsApp stores all message data on the user's device in an encrypted SQLite database. We used a rooted Android device to obtain the decryption key and

[3]For brevity, we refer to COVID-19 simply as COVID and use these terms interchangably

and extracted the decrypted database every week. The media content, which is stored as encrypted URLs was downloaded locally and decrypted using a public tool[4] (slightly modified for our convenience). WhatsApp deletes media content from their servers after a certain amount of time. As a result, when decrypting media files, we missed a small subset of the content shared (14%). The joining of the groups took place over a 1 month period, as new groups were being identified. The data collection started from 10 January 2020 onward and continued until 23 February 2020. We have complete data from all groups from the end of February until the second week of April. The details of the dataset are summarized in Table I.

Table I: Overview of our WhatsApp dataset.

#Groups	227	
#Admins	521	
#Users	18,475	
#Unique users	16,493	
Total #messages	60,202	
#Text messages	28,497	47%
#Images	14,633	24.5%
#Video	11,196	18.6%
#Audio	2,688	4.5%
Others	6,740	5.4%
#URLs	3,188	2.5%

Twitter data. To compare the data we obtained from the WhatsApp groups to other open, well studied social media platforms, we also gathered data from Twitter. Specifically, we obtained historical Twitter data on an extensive list of hashtags specific to COVID-19 in Pakistan such as #CovidPakistan, #CoronaFreePakistan[5] and other local Twitter trend variations. This gave us 800,000 tweets.

Ethics note. The groups joined, had been openly propagated on Facebook, Twitter, and other mediums and can be joined by anyone. The profile bio of our WhatsApp account declares that we are collecting information for research purposes. We also anonymized the user data, before analyzing it.

B. Identifying COVID-19 Text Messages

We extract COVID-19 related text messages using a keyword-filtering approach. We utilize [24], which offers a dictionary of COVID-19 *English* keywords. We added small variations and multiple spellings to the dictionary to capture a wide variety of content related to the pandemic. We translated these keywords into *Urdu* and used both the English and Urdu keywords to search our dataset. The final list includes keywords such as "corona", "coronavirus", "covid-19", "covid", "covid19", and "corona virus", among others. This keyword based approach results in a high precision yet low recall method to identify COVID related messages. Using this approach, we obtained 5,039 COVID related text messages between March 16, 2020 and April 09, 2020. Figure 1 compares the number of daily COVID-19 related and Non-COVID-19 related text messages in our dataset.

[4]https://github.com/ddz/whatsapp-media-decrypt
[5]https://cutt.ly/nyXrVyp

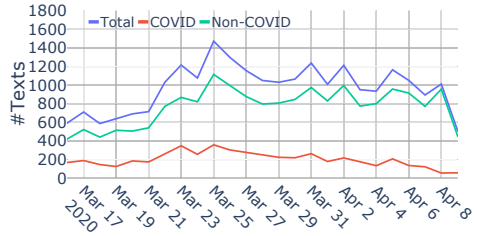

Figure 1: COVID vs. Non-COVID texts. Timeline of number of messages and messages containing COVID related keywords in our WhatsApp dataset

C. Identifying COVID-19 Images

As we see in Table I, around 25% of the content is images. Hence, solely evaluating text would give a distorted view of the overall information landscape. Naturally, image content is far harder to automatically categorize. Therefore, to extract images discussing COVID-19, manual tagging was performed

Two annotators tagged a total of 6,699 images, ranging from 16 March to 9 April, 2020. An image was declared as COVID related if it had any of the following attributes:

1) Contained Coronavirus, COVID-19, or any other related terminology in Urdu or English.
2) Information relating to a lockdown or any restrictions being imposed/relaxed by the government on business or public/private institutions.
3) Sharing of any precautionary measures like prayers for protection from disease, herbal medications, etc.
4) Contained any references to the environmental or economic impact of COVID-19.
5) Contained people with personal protective equipment, possible quarantine centers, and people practicing or encouraging social distancing.

An inter annotator agreement score of 98% was observed. In cases of a conflict, the annotators were allowed to mutually discuss and agree upon a label. A total of 2,309 (34.5%) images were identified as COVID related, while 4,390 (65%) were identified as non-COVID images. For context, Figure 2 shows the percentage of COVID related images over time. We see that, as the pandemic intensifies, so does the fraction of related images.

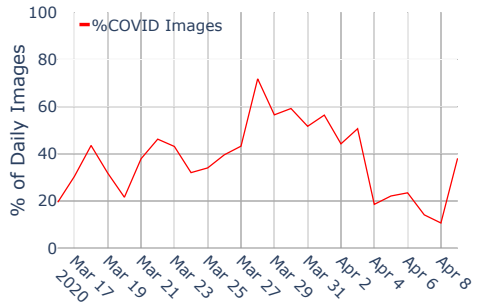

Figure 2: Timeline of the percentage of images and images containing COVID related content in our WhatsApp dataset

IV. RQ1: Information Sharing on COVID-19

In this section, we answer our first research question, to understand what types of COVID related information is being shared on WhatsApp. We first annotate the data into 5 overalapping categories and then use these categories to understand the types of information.

A. Message Type Categorization

In order to further characterize the types of COVID-19 related content being shared within our WhatsApp groups, we categorized the COVID related text and images manually into the five categories: *Information*, *Misinformation*, *Jokes/Satire*, *Religious*, and *Ambiguous*. The categories were chosen based on a preliminary manual exploration of the COVID-19 content. The categories are not mutually exclusive and as a result, a single message (text/image) can belong to multiple categories. We describe each category below:

1) **Information:** This category consists of WhatsApp content that contains either factual News or COVID related facts. News reports are fact checked using Poynter's COVID Facts database[6] which contains all of the falsehoods detected by a large number of fact checking organizations. In addition, AFP Pakistan Fact Check[7] is used to verify news articles. The contents of the text or image are evaluated against the falsehoods in the database to verify their validity. Google search was also used to verify certain claims not present in the Poynter dataset. If the news is reported by a reputed news source, then it is labelled as "Information". A news source is considered reputed if it has a satellite news channel or newspaper at a national level. COVID related facts are verified using WHO's COVID Information and pervalent myths.[8]
2) **Misinformation:** This category is the inverse of the above 'Information' category. Any content which is either verified to be misinformation or could not be verified as credible information is placed in this category. Content was checked using Poynter COVID-Facts and Falsehoods database, AFP Pakistan and WHO's COVID Informations and COVID Myths.
3) **Jokes/Satire:** This class consists of content that intends to poke fun at the COVID-19 pandemic itself or any COVID related government/political actions using sarcasm, satire or memes. It also contains content that consists of non-factual opinions/analysis regarding current COVID related events or government actions.
4) **Religious:** Since Pakistan has a 98% Muslim population, religion plays an important role in information dissemination. A religious theme in content is identified by looking for (i) references to spiritual texts, (ii) quotes of religious scholars (called *Maulana*, *Mufti*, or *Sheikh*), and (iii) mentions of religious acts such as prayer, fasting etc.
5) **Ambiguous:** If the content does not have enough information to be classified into one or more of the above categories, it is then assigned to the 'Ambiguous' category. This category mainly consists of content where people are distributing Personnel Protective Equipment (PPE), social media requests to follow/subscribe, contact information of NGO's, donation requests, etc.

[6]https://www.poynter.org/ifcn-covid-19-misinformation/
[7]https://factcheck.afp.com/afp-pakistan
[8]https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters

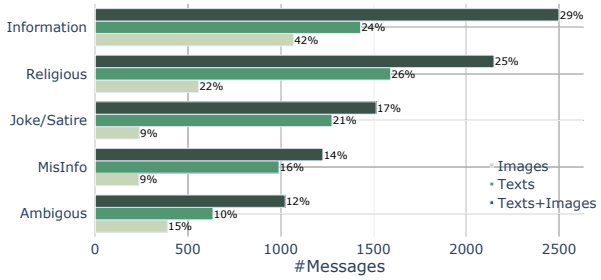

Figure 3: Percentage of COVID-19 images (Light Green), texts (Medium Green) and texts + images (Dark Green) for each category. Notably, 14% of the total messages were labeled to be misinformation.

Note that to maintain consistency in annotation quality between images and text (images were annotated by two annotators whereas text was done by only one annotator), we randomly sampled 25% of the 5,039 COVID-19 related texts, and validated the annotation. We showed the 25% random sample of messages to one additional annotator and measured the agreement with the original annotator. We find an 82% agreement between the two annotators, with one or more common labels counted as an agreement for our non-mutually exclusive classes. The majority of disagreements were between Information, Jokes/Satire and Religious classes. This is because a lot of texts contain different proportions of the three. Very few disagreements were observed when one of the annotators tagged a text as Misinformation, which were resolved after a discussion between the two annotators.

B. Message Type Analysis

We now analyze the different types of COVID-19 related content, in both texts and images on WhatsApp, based on the above annotation. We have a total of 5,039 texts and 2,309 images which discuss COVID related information between March 16 and April 9, 2020. The overall distribution of texts and images into the COVID-19 related content categories are shown in Figure 3. A majority of the content is simple information (29%), containing news articles, latest government actions and health information related to COVID-19. This is followed by religious content (25%). The large amount of religious content emphasizes the importance of religious sentiment within the society, especially during times of uncertainty created by the pandemic. Religious scholars and Holy Verses from religious books were cited in these messages. Religious content was also event focused. For instance, a ban on congregational prayers and the rigorous COVID testing of a group of religious people on a proselytising trip resulted in messages criticising these government actions.

This was followed by Jokes/Satire representing 17% of the messages. Political actions by rival parties, and government officials were frequently ridiculed and mocked, including personally targeted attacks. For instance, many

of such texts were against government action of opening the border with Iran, blaming the officials for bringing Corona to Pakistan. Interestingly, a non trivial amount (14%) was Misinformation, which shows that one in seven messages shared contained some misleading information. This included fake news reporting deaths of politicians, fake quotes from famous personalities and international figures, or fake COVID origin stories. We show detailed analysis of misinformation in Section V.

Finally, a small fraction of messages (12%) were labelled "Ambigious". Among these, a significant portion of texts contained Facebook and YouTube follow/subscribe requests to COVID related pages and channels, donation requests, or shared contact information for COVID-affected and poverty-stricken people. We also found some images depicting quarantine centers, hospitals, doctors, and patients, which did not fit into the above categories due to the lack of context.

C. Lifetime of Messages

In this section, we try to understand the *impact* of the various types of COVID messages on WhatsApp. To do so, we analyzed the lifetime of various types of COVID related messages. The lifetime of a message is the difference between the last and first appearance of a message (in hours) in our dataset. First, we grouped together perceptually similar images using a popular, state of the art image hashing tool from Facebook known as PDQ hashing.[9] The hashes were generated for all the COVID related images and then instances of similar images were clustered together by using Hamming distance, with a threshold of 70%, between the hashes. The difference between first and last appearance of a representative image in each cluster was considered as the lifetime of that image. For texts, exact string matching was used to find the first and last appearance of a text. Table II shows the mean lifetime of messages belonging to the various COVID-19 content categories.

Each category exhibits distinct mean and variance measure for lifetime. The most short lived messages belong to the "Jokes/Satire" category. This appears logically coherent since jokes, opinions and satirical texts are generally dictated by events and die out quickly as the public focus shifts from one event to another. Interestingly, the lifetime of a message containing misinformation is quite high, for both text (7 hrs) and images (5.5 hrs), especially compared to the Information category. This means that misinformation tends to persist longer compared to information, which supports existing studies showing similar results [25], [14]. Given that WhatsApp is a closed platform with no content moderation or third party fact checking, the fact that misinformation tends to stick around longer might be expected, but might also be problematic when compared to other social networks, where eventually corrections can be issued.

V. RQ2: COVID-19 Misinformation

In this section, we specifically look at misinformation posts, and characterize the types of misinformation shared on WhatsApp. We first categorize the types of misinformation based on reports from popular fact checking organizations and then base our characterization on this categorization.

[9]https://github.com/facebook/ThreatExchange

Table II: Lifetimes of COVID-19 related texts and images shared on WhatsApp. *Misinformation tends to have the highest mean lifetime.*

Text Messages			
Label	**Num. texts**	**Mean (hrs)**	**Std Dev (hrs)**
Information	1108	2.75	21.98
Religious	829	6.98	29.15
Jokes/Satire	919	1.92	9.15
Misinformation	596	7.0	28.03
Ambiguous	313	10.05	39.2
Images			
Label	**Num. images**	**Mean (hrs)**	**Std Dev (hrs)**
Information	1069	0.55	2.87
Religious	557	2.70	6.14
Jokes/Satire	238	1.21	4.07
Misinformation	236	5.57	9.17
Ambiguous	389	1.35	4.31

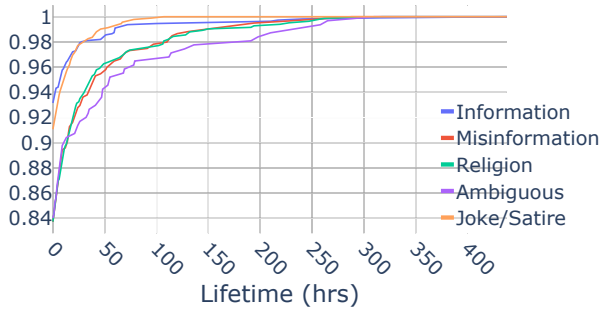

Figure 4: CDFs of Text of COVID-19 categories (Note the broken y-axis: For better Interpretability).

A. Misinformation Message Analysis

The distribution of the various categories is shown in Figure 5.

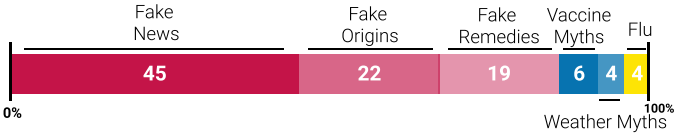

Figure 5: Percentage of texts on WhatsApp for each type of COVID-19 related misinformation.

Fake News. The most frequent form of COVID related misinformation is in the form of fake news with 45% of misinformation texts. This includes fake news pertaining to COVID positive tests and COVID related deaths of world figures such as Ivanka Trump, Prince Williams and even the current Prime Minister of Pakistan, Imran Khan. Conspiracy theories about Bill Gates intending to place RFID chips in people to track COVID-19 were also seen. Ironically, fake news were also observed regarding a doctored government action announcing 'Punishment for Spreading Fake News on social media'.

Fake Origins. The second most prevalent form of COVID-19 related misinformation is claiming fake origin stories for the virus with 22% of the misinformation texts. Fake origin stores include a Corona named lake in Kazakhstan from which the virus came to being grown in a lab in

China or United States. A few Hollywood movies, namely 'Contagion', 'Resident Evil' and 'I am Legend' along with the book 'The Eye of Darkness' were frequently mentioned stating that COVID-19 had been predicted by them.

Fake Remedies. Making up roughly 20% of the misinformation, this type contains bogus remedies and treatments such as the 1-minute breath hold test to detect COVID, and various items like basil seeds, gargling with salt or garlic water, honey lemon tea and even Hepatitis-C medicine as cures to COVID-19.

Vaccine Myths. The fake origin stories were sometimes accompanied with claims of the vaccine already being developed and being used as an economic leverage. Countries such as Israel, China and United States were mentioned with claims of the vaccine already developed. This category makes up around 6% of the misinformation texts.

Weather Myths. Four percent of the misinformation claims that the virus can not survive in winter, summer or rainy seasons and the outbreak would die down on its own.

Flu Comparison. Only 2% of the misinformation attempted to downplay the symptoms and severity of the disease by comparison to the common seasonal flu. Even though this narrative was popular elsewhere (e.g. US), it did not have much salience in Pakistan, with the general public acknowledging COVID-19 as a distinct and more dangerous disease as compared to the common flu.

B. Lifetime of Misinformation

The temporal properties of the various categories of misinformation are analyzed in table III. For each message, we compute the lifetime as the difference between its last and first occurance. The 'Fake News' category has the shortest lifespan as evidenced by the lowest mean of 4 hrs. This seems to be consistent with the hypothesis that event triggered content is short-lived, with similar properties to the 'Jokes/Satire' category of COVID-19 related textual content. The highest lifespan is for the 'Fake Remedies' category with a mean life of 10 hrs, which is significantly larger than other major categories. This indicates that content that is not tethered to a social event is more likely to being in circulation on a social media platform like WhatsApp.

Table III: Lifetime of misinformation texts shared on WhatsApp.

Label	Num. texts	Mean (hrs)	Std Dev (hrs)
Fake News	307	4.06	18.7
Fake Origins	171	9.4	35.3
Fake Remedies	125	10.6	33.8
Weather Myths	26	27.57	67.6
Flu Comparison	16	6.68	16.66

VI. RQ3:User Behaviour Analysis

Every WhatsApp group has two types of users: (1) *producers* and (2) *consumers*. Some users share and post messages whereas others silently observe. In general, producers are few and consumers are many (Table I). In this section, by examining the user behavior, we hope to understand if there is any deliberate spread of disinformation.

We use "UpSet"[10] plots, to visualize user behavior, where every set is a unique user. The bottom matrix (combination matrix) of Figure 6,7 shows the intersections of the sets across COVID categories, while the bars on the top indicate the number of users (sets) within that intersection. The bars on the left indicate total users (sets) within a given category.

A. Text Sharing Trends

The UpSet plot in Figure 6 is plotted against the text messages shared by individual users. If we observe the combination matrix, we observe that users are sharing textual content belonging to single category. The most exclusively shared category is "Ambigous". It can be attributed to the users that join WhatsApp groups intending to share advertisements and call-for-donations only. The second and third highest intersection sets are for "Religious" and "Misinformation" being shared exclusively. This deviates from the trend observed for Images. People are more likely to **exclusively** share "Texts" containing "Misinformation" as compared to "Images" containing "Misinformation". This prompts for need for more research in finding traces of disinformation within text messages.

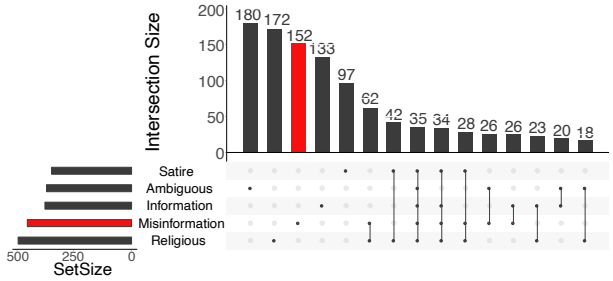

Figure 6: UpSet plot for users posting COVID related texts. Top 15 intersection sets are visualized. *More users appear to share texts that belong to a single category.*

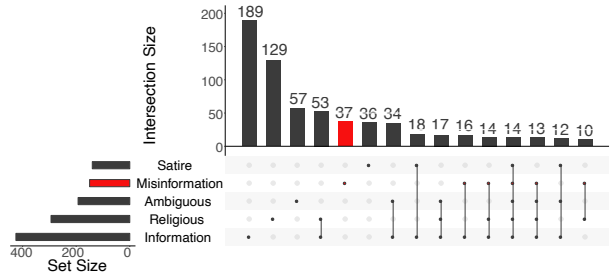

Figure 7: UpSet plot for users posting COVID related images. Only the top 15 intersection sets are visualized. *A lot of users are sharing information and religious content, whereas some share misinformation.*

B. Image Sharing Trends

The UpSet plot shown in Figure 7 is made against the type of Images individual users are sharing. It is good to see that the majority of users are sharing correct information about the pandemic. An encouraging trend is that users

[10]For an introduction, see https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4720993/.

are not exclusively sharing misinformation, rather a mix of content is being shared. Only 37 users exclusively shared misinformation, whereas 67 users shared a mix of content, along with misinformation.

To further understand if there is disinformation, we tried to see if a specific type of image is being spread more than others. Using the clusters we already had, made via PDQ hashing and Hamming distance, clusters having misinformation were identified. The number of images within these clusters are a good indicator of the impact of given misinformation on the network. 23 unique images were shared more than 1 times and only 8 were shared more than 5 times.

This implies that even if we consider the images shared multiple times to be disinformation, the quantity of disinformation is very low. Hence we believe, the misinformation is, rather than being an organized effort, is mainly being spread due to lack of awareness.

VII. RQ4: Cross Network Information Spread

Finally, we answer how information flows between WhatsApp and Twitter. Given that each social network has different properties (closed vs. open), affordances (e.g.,the ability to see how popular a content is with retweet/like count) and user bases, such a comparison is interesting.

A. Methodology

Twitter was chosen as most of Twitter data is public, and it serves as one of the major information conduits. We obtained more than 0.8 million unique tweets starting from January 10 to April 9 using an exhaustive list of hashtags related to COVID-19 in Pakistan. It must be noted, that both Twitter and WhatsApp datasets only represent a subset of the actual activity and in no way can be thought to represent the full networks' behaviour. In order to understand the flow of information, images and text present within WhatsApp, from 16 March to 9 April 2020, were compared with tweets in the same time range. To find similar content across Twitter and WhatsApp, PDQ hashing was used for images and fuzzy string matching was used for text messages.

B. Cross Platform Image Spread

To understand how images were propagated across networks, we isolate the tweets containing image content from the Twitter dataset. This covers a total of 67,119 images. We then generate PDQ hashes for both WhatsApp and Twitter images, and matching two images if their hashes have a Hamming distance of 40 (default value suggested by PDQ). Around 1,500 similar images were found common to both WhatsApp and Twitter, within the date range of interest (16 March and 9 April). Out of these 1,500 images, 541 were COVID-19 related.

Table IV breaks down the images into the categories previously described, alongside the average number of retweets, replies and the lifetime of the image (difference between first and last appearance of an image on Twitter). Firstly, we observe that the largest category of images shared across both Twitter and WhatsApp is that of misinformation (29%). We can also see that misinformation tweets have a high average number of retweets, potentially reaching tens of thousands of users. This is signal in Twitter (or other social networks) which does not exist on WhatsApp, where the social popularity signals like retweet or like counts are shown.

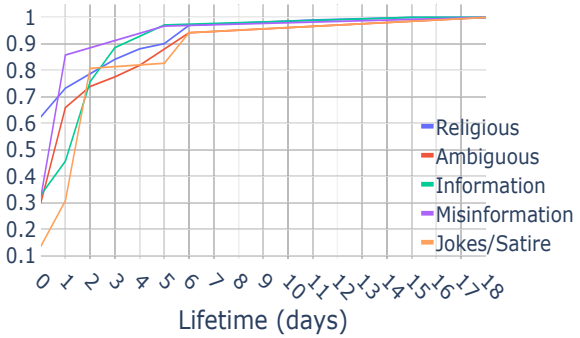

Figure 8: CDFs of life of an image, along with content type, as seen on Twitter. Twitter tends to hold a message alive for a coupe of days. *A healthy trend is that Information tends to live the longest on Twitter.*

Compared to the other categories, misinformation on Twitter tends to die quicker. Figure 8 shows a CDF of the lifespan of the various categories of images on Twitter. We see that most types last at most a day, with most having a long tail lasting weeks. This is very different from the pattern we observed on WhatsApp, where on average, most images lasted only a few hours. Images containing information have a much higher lifetime, which is also in stark contrast to what we observed on WhatsApp.

The difference in longevity of a message, as seen in Table II and IV, points towards the nature of interactions. Interactions on WhatsApp are immediate, as new messages constantly replace old ones. Whereas on Twitter old tweets can be easily brought back into limelight, using retweets, comments and likes by influential people. Hence the ability of Twitter to keep conversation, around a tweet, alive for a long time could be the reason for the overall life of COVID tweets (in days) compared to WhatsApp (in hours).

Table IV: Characteristics of images mapped between Twitter and WhatsApp.

Label	**Num. images**	**Retweets** (Mean)	**Replies** (Mean)	**Life** (days)
Information	79	64.18	75.27	5.05
Religious	108	42.66	45.82	3.25
Jokes/Sarcasm	104	26.96	89.71	3.0
Misinformation	183	67.72	444.28	1.6
Ambiguous	146	156.82	309.64	4.2

A schematic diagram of the temporal flow of images across WhatsApp and Twitter is shown in Figure 9. The Figure also provides three example case studies of images which originated on WhatsApp and went on to become widely retweeted on Twitter. From our analysis of the timelines of images observed on both platforms, we can conclude that most of the images are seen originating from WhatsApp and then appearing on Twitter (in our dataset). On average an image appears on WhatsApp 4 days earlier as compared to Twitter. As a result in light of the data analyzed, it can be concluded that WhatsApp plays a critical role in COVID related content dissemination to other networks

in Pakistan. This is especially important in the context of results from Table IV with a majority (29%) of the content that is common between the two platforms being misinformation, compared to only 12% being information.

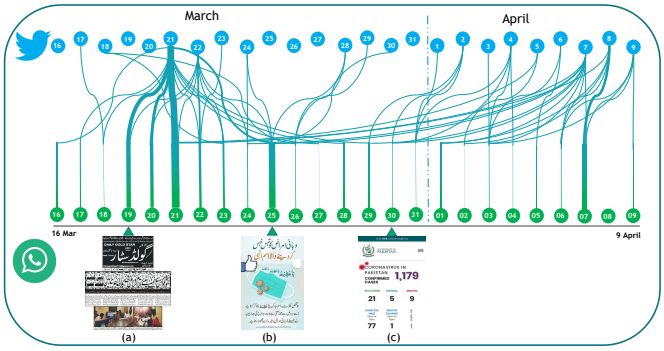

Figure 9: COVID images' temporal flow across WhatsApp and Twitter (a line's thickness depicts the number of images flowing across). *Some Observations*: *a)* a news snippet originates from WhatsApp on 19th March and is seen on Twitter on 21st; *b)* religious supplication to fight COVID is observed on WhatsApp 2 days earlier than on Twitter; *c)* official stats of COVID patients seen on 30th March on WhatsApp earlier than on Twitter.

VIII. Conclusions

In this paper, we have provided the first detailed analysis of Pakistani WhatsApp public groups, focusing on the COVID-19 discourse. We have analyzed WhatsApp text and image messages collected for more than 6 weeks from 227 public WhatsApp groups to shed light on the salient misinformation dissemination trends and to share insights on how Pakistani social media users are experiencing and responding to the COVID-19 pandemic. Our work is unique as this is the first work to not only study misinformation trends on WhatsApp but also find a relation between WhatsApp and Twitter. Our analyses showed that while it is true that the majority of shared information is not misinformation, misinformation seems to have a longer lifespan on WhatsApp compared to other types of COVID messages (the lifetime of misinformation is roughly 4 times that of correct information). On Twitter the inverse was seen, as COVID misinformation tended to disappear from Twitter 3 times faster than information. This can potentially be attributed to the open nature of Twitter, and how a vast number of users can publicly negate such tweets. While observing user behavior, we found 8 images that could be attributed to organized disinformation, other than that, we did not find any evidence of disinformation within images. Whereas more work is required in detecting disinformation via text messages. We conclude by saying that our dataset has only scratched the surface of how user interactions happen on WhatsApp. More work needs to be performed to understand user behavior, and new ways need to be proposed to detect misinformation in such closed networks.

References

[1] "Two Billion Users–Connecting the World Privately," dated: February 12 2020. [Online]. Available: https://tinyurl.com/ybwq4hpa

[2] G. Resende, P. F. Melo, H. Sousa, J. Messias, M. Vasconcelos, J. M. Almeida, and F. Benevenuto, "(Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures," in *WWW '19*, 2019.

[3] G. Resende, P. Melo, J. C. S. Reis, M. Vasconcelos, J. M. Almeida, and F. Benevenuto, "Analyzing Textual (Mis)Information Shared in WhatsApp Groups," in *Proceedings of the 10th ACM Conference on Web Science*, 2019.

[4] N. Newman, R. Fletcher, A. Kalogeropoulos, and R. K. Nielsen, "Reuters Institute Digital News Report 2019 ," Reuters Institute for the Study of Journalism, 2019.

[5] "WHO Director-General's media briefing on COVID-19," dated: March 11 2020. [Online]. Available: https://tinyurl.com/WHO-DG-OpeningRemarks11March20

[6] A. Boadle, "Facebook's whatsapp flooded with fake news in brazil election," 2018. [Online]. Available: https://www.reuters.com/article/us-brazil-election-whatsapp-explainer

[7] B. Perrigo, "How Volunteers for India's Ruling Party Are Using WhatsApp to Fuel Fake News Ahead of Elections," 2019. [Online]. Available: https://time.com/5512032/whatsapp-india-election-2019/

[8] V. Goel, "In India, Facebook's WhatsApp Plays Central Role in Elections," May 2018. [Online]. Available: https://www.nytimes.com/2018/05/14/technology/whatsapp-india-elections.html

[9] C. Lokniti, "How widespread is WhatsApp's usage in India?" 2018. [Online]. Available: https://www.livemint.com/Technology/O6DLmIibCCV5luEG9XuJWL/How-widespread-is-WhatsApps-usage-in-India.html

[10] K. Garimella and G. Tyson, "Whatsapp, doc? a first look at whatsapp public group data," in *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[11] R. Evangelista and F. Bruno, "Whatsapp and political instability in brazil: targeted messages and political radicalisation," *Internet Policy Review*, vol. 8, no. 4, pp. 1–23, 2019.

[12] G. Resende, P. Melo, H. Sousa, J. Messias, M. Vasconcelos, J. Almeida, and F. Benevenuto, "(mis) information dissemination in whatsapp: Gathering, analyzing and countermeasures," in *The World Wide Web Conference*, 2019, pp. 818–828.

[13] A. Yadav, A. Garg, A. Aglawe, A. Agarwal, and V. Srivastava, "Understanding the political inclination of whatsapp chats," in *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, 2020, pp. 361–362.

[14] K. Garimella and D. Eckles, "Images and Misinformation in Political Groups: Evidence from WhatsApp in India," *arXiv preprint arXiv:2005.09784*, 2020.

[15] P. de Freitas Melo, C. C. Vieira, K. Garimella, P. O. S. V. de Melo, and F. Benevenuto, "Can WhatsApp Counter Misinformation by Limiting Message Forwarding?" 2019.

[16] A. Maros, J. Almeida, F. Benevenuto, and M. Vasconcelos, "Analyzing the use of audio messages in whatsapp groups."

[17] N. Purnell, "Facebook's whatsapp battles coronavirus misinformation," April 2020. [Online]. Available: https://www.wsj.com/articles/facebooks-whatsapp-battles-coronavirus-misinformation-11586256870

[18] F. Jin, W. Wang, L. Zhao, E. Dougherty, Y. Cao, C.-T. Lu, and N. Ramakrishnan, "Misinformation propagation in the age of twitter," *Computer*, no. 12, pp. 90–94, 2014.

[19] I. C.-H. Fung, C. H. Duke, K. C. Finch, K. R. Snook, P.-L. Tseng, A. C. Hernandez, M. Gambhir, K.-W. Fu, and Z. T. H. Tse, "Ebola virus disease and social media: a systematic review," *American journal of infection control*, vol. 44, no. 12, pp. 1660–1671, 2016.

[20] K. Sharma, S. Seo, C. Meng, S. Rambhatla, A. Dua, and Y. Liu, "Coronavirus on social media: Analyzing misinformation in twitter conversations," *arXiv preprint arXiv:2003.12309*, 2020.

[21] L. Singh, S. Bansal, L. Bode, C. Budak, G. Chi, K. Kawintiranon, C. Padden, R. Vanarsdall, E. Vraga, and Y. Wang, "A first look at covid-19 information and misinformation sharing on twitter," *arXiv preprint arXiv:2003.13907*, 2020.

[22] R. Kouzy, J. Abi Jaoude, A. Kraitem, M. B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E. W. Akl, and K. Baddour, "Coronavirus goes viral: Quantifying the covid-19 misinformation epidemic on twitter," *Cureus*, vol. 12, no. 3, 2020.

[23] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala, "The COVID-19 social media infodemic," *arXiv preprint arXiv:2003.05004*, 2020.

[24] S. K. Rashed, J. Frid, and S. Aits, "English dictionaries, gold and silver standard corpora for biomedical natural language processing related to sars-cov-2 and covid-19," 2020.

[25] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.