Social Network Analysis and Mining (2022) 12:5
https://doi.org/10.1007/s13278-021-00833-0

ORIGINAL ARTICLE

A deep dive into COVID-19-related messages on WhatsApp in Pakistan

R. Tallal Javed[1] · Muhammad Usama[1,3] · Waleed Iqbal[4] · Junaid Qadir[1,2] · Gareth Tyson[4] · Ignacio Castro[4] · Kiran Garimella[5]

Received: 21 June 2021 / Revised: 2 October 2021 / Accepted: 5 October 2021 / Published online: 15 November 2021

Abstract
The spread of COVID-19 and the lockdowns that followed led to an increase in activity on online social networks. This has resulted in users sharing unfiltered and unreliable information on social networks like WhatsApp, Twitter, Facebook, etc. In this work, we give an extended overview of how Pakistan's population used public WhatsApp groups for sharing information related to the pandemic. Our work is based on a major effort to annotate thousands of text and image-based messages. We explore how information propagates across WhatsApp and the user behavior around it. Specifically, we look at political polarization and its impact on how users from different political parties shared COVID-19-related content. We also try to understand information dissemination across different social networks—Twitter and WhatsApp—in Pakistan and find that there is no significant bot involvement in spreading misinformation about the pandemic.

Keywords Misinformation · Infodemic · Social computing

1 Introduction

Applications like Twitter, Facebook, and WhatsApp are enabling millions of users to connect and interact. This has led to sharing ideas, getting exposed to different ideologies, and absorbing information at an unprecedented pace. Out of all these services and apps, WhatsApp is the most popular and widely used medium of communication. This has created a closed social network of more than 1.5 billion users. Apart from having a huge user base, it also has the most active users at a time, out of all the social networks (Two Billion Users 2020).

This makes WhatsApp a very important medium for analysis, as it is a major tool for opinion formation and social exchange. Due to its end-to-end encryption, it is becoming a medium of choice for anti-government movements, sharing of radical ideas, and gang operations (UK says WhatsApp 2017). Similarly, WhatsApp is also used for the propagation of antisocial behavior. A study on Brazil's WhatsApp users (Resende et al. 2019a) revealed how WhatsApp can be an effective tool for the spread of disinformation. A study conducted in India (Saha et al. 2021) revealed the spread of hate speech and Islamophobia on WhatsApp. These problems are exacerbated by the fact that content moderation in WhatsApp is rather limited. The content of a group is only moderated by the groups' administrators. Admins have very few

✉ R. Tallal Javed
tallal.javed@itu.edu.pk

Muhammad Usama
muhammadusama@lums.edu.pk

Waleed Iqbal
w.iqbal@qmul.ac.uk

Junaid Qadir
jqadir@qu.edu.pk

Gareth Tyson
g.tyson@qmul.ac.uk

Ignacio Castro
i.castro@qmul.ac.uk

Kiran Garimella
garimell@mit.edu

1 Information Technology University of the Punjab, Lahore, Pakistan

2 Qatar University, Doha, Qatar

3 Lahore University of Management Sciences, Lahore, Pakistan

4 Queen Merry University, London, UK

5 MIT, Cambridge, USA

tools at hand: either restrict who can post content or remove certain users from the group. Group admins cannot even do the simple moderation task of deleting a user's post. As a result, the moderation abilities on WhatsApp are very scant.

This work is an extended version of our earlier preliminary analysis of COVID-19-related messages being disseminated across public WhatsApp groups by Pakistani users (Javed et al. 2020). In this extended version, in addition to looking at the type of messages being disseminated across COVID-19 in Pakistani WhatsApp groups, we also try to understand the impact that political affiliation has on a group's overall sentiment related to COVID-19 and the types of messages being propagated. Our main research questions are:

RQ1: What type of messages about COVID-19 are disseminated in public WhatsApp groups of Pakistan?
RQ2: What is the general user behavior when sharing a message? Specifically, is there a connection between a group's political affiliation and the content that is being shared?
RQ3: Is there reciprocation between information dissemination related to COVID-19 over WhatsApp and Twitter?
RQ4: What type of sentiment is expressed by users when sharing COVID-19 messages? Does this vary on the basis of political affiliation?

To answer these research questions, we have gathered the data from 227 publicly accessible WhatsApp groups during January 10, 2020–April 9, 2020. This dataset is the first of its kind, giving us a unique opportunity to analyze COVID-19-related discussion from one of the largest countries in the world involving a multi-modal environment—images and text—and multiple social networking platforms of communication—WhatsApp and Twitter.

We start by analyzing the content disseminated in these 227 public groups and extract the COVID-19-related content out of them. Extracted content is separated into image, video, text,[1] documents and links. Using our dataset which we make publicly available,[2] we make the following contributions:

- We analyze the groups and messages therein with a focus on political affiliations.
- We give an overview of how users with different political affinities spread COVID-19-related messages. We also show that, on average, 14% of the content shared is misinformation.
- We give an overview of the prevailing sentiment of COVID-19 posts and find that the sentiment is mostly negative.

2 Related work

2.1 Misinformation on social media

Many studies have been conducted that gave a special focus to rumors and misinformation prior to the events of the 2016 US presidential elections (Starbird et al. 2014). However, the events of the US presidential election have since triggered a flurry of work on this topic. As a result, there have been many papers attempting to understand the impact social media has, the amount of misinformation present on it and the amount of exposure users have to this information (David et al. 2018; Iosifidis and Nicoli 2020; Bovet and Makse 2019). For example, during the 2016 elections, social media were used extensively to manipulate social media users to sway their political inclinations. Grinberg et al. (2019) analyzed Twitter to understand the extent of political manipulation present during this period. Similarly, Badawy et al. (2019) used Twitter to understand the effects the Russian Internet Research Agency might have had on American Twitter users. They characterize the interactions of Twitter Republican and Democratic users with the Russian trolls.

There are various methods for detecting misinformation. They can be divided into two major approaches: content-based and propagation-based. Habib et al. (2019) performed a systematic literature review to understand different methodologies for detecting misinformation in online social networks. Chen et al. (2020) performed an analysis of Twitter to understand different types of misinformation. Their analysis relied on a graph of users based on the content they share. In other similar works (Zollo and Quattrociocchi 2018; Cinelli et al. 2019), the authors provide analysis on social media users and their interactions with controversial topics and content. A study on Instagram was conducted by Trevisan et al. (2019) where they gave detailed insights on how users interact with political content. Similarly, many researchers have conducted independent studies in line with those mentioned earlier (Zhang and Ghorbani 2020; Shu et al. 2020, 2017; Sharma et al. 2019; Zhou and Zafarani 2018; Zarei et al. 2020).

2.2 Analyzing content on WhatsApp

The WhatsApp messaging service is the most actively used online social network in the world. WhatsApp is a closed

[1] In this study, we only focus on text and images and leave the analysis over video content for future work.

[2] https://tinyurl.com/snam-pakistan.

network, without any official access for analyzing its content. As a result, not a lot of work has been done to analyze its content and user interactions. It is a documented fact that WhatsApp is being actively used for the dissemination of misinformation (Boadle 2018; Perrigo 2019). Due to the popularity of WhatsApp in certain regions, political parties have been actively using WhatsApp groups to reach the masses (Goel 2018). Surveys performed in Brazil and India (Lokniti 2018; Newman et al. 2019) (two of the largest democracies) show that one in six users are part of a political group on WhatsApp.

Garimella and Tyson (2018) were the first to provide tools for analyzing public WhatsApp groups and collecting data at scale. These tools have enabled researchers to study WhatsApp at scale and created a window into the world of WhatsApp. Some of the recent studies can be found at Evangelista and Bruno (2019), Resende et al. (2019b), Yadav et al. (2020), Garimella and Eckles (2020), where researchers analyzed public WhatsApp groups in various contexts. Resende et al. (2018) gave a unique insight by analyzing doctored images used to fuel political smear campaigns against opposing parties on public WhatsApp groups in Brazil. Similarly, Garimella and Eckles (2020) analyze the images in Indian WhatsApp groups during the 2019 Indian elections. The study found that 13% of the images contained misinformation, using reverse image search on Google images. In parallel Melo et al. (2019) gather, analyze, and visualize public WhatsApp groups and identify the extent of misinformation found in India, Indonesia, and Brazil. Apart from text messages and images, WhatsApp has a large content of audio files. Maros et al. (2020) analyzed 330 public WhatsApp groups and proposed that audio messages containing misinformation spread much more farther and wider.

2.3 The COVID-19 infodemic

Our work revolves around understanding health information being shared on WhatsApp, and how users interact with it while considering the political inclination of users. The COVID-19 pandemic has also created an infodemic, as declared by the World Health Organization. An infodemic refers to the inflow of information that is so large that users are unable to discriminate effectively between misinformation and correct information. It is already documented that WhatsApp is a source of misinformation related to health (Purnell 2020), ranging from wrong symptoms to ineffective treatments (Bhatnagar and Choubey 2021; Javed et al. 2020). This makes it critical to understand the health content present in WhatsApp groups.

Apart from WhatsApp, researchers have provided a dashboard to analyze health misinformation on Twitter (Sharma et al. 2020). They analyze 25 million tweets and also perform a country-wide analysis of sentiment. They also provide an up-to-date view of how people are reacting to COVID-19-related content on Twitter. Similarly, Singh et al. (2020) look at Twitter based misinformation about COVID-19 and give insights about the propagation of misinformation on online social networks is in line with the rise of cases in a given demographic. Another study on Twitter (Kouzy et al. 2020) found that some tags have more misinformation than others, pointing toward potential safe tags on Twitter. Cinelli et al. (2020) analyze different social networks for COVID-19-related content. They analyze, Twitter, Instagram, Reddit, Gab, and YouTube, giving a comprehensive picture of the state of COVID-19 content on these websites. They not only do content analysis but also try to understand the propagation of misinformation on these social networks.

2.4 Our work's novelty

This work is an extension of our previous work (Javed et al. 2020). In this study, we analyze COVID-19-related discussions on WhatsApp and explore the political influence in this context. Since WhatsApp is a popular and frequently used application, it is critical to understand how the populace is utilizing the platform during the pandemic and how the platform facilitates the spread of misinformation. While doing so, we extract valuable insights from the dataset and try to quantify how misinformation and politics can be intertwined.

Studies have been conducted to analyze WhatsApp messages for political events in Brazil and India. On the contrary, this study tries to understand the impact WhatsApp is having on the infodemic in Pakistan, while considering the political nature of our groups. Since Pakistan is a Muslim majority country, religion is relevant in the daily life of the Pakistani citizens,[3] we offer a first insight into how religion and politics together play a role in this infodemic.

3 Methodology

We use a multitude of techniques for data collection, annotation, and analysis of data. We also create novel algorithms to better understand the content being shared. The details of these methods are discussed below.

3.1 Dataset preparation

WhatsApp allows its user base to create either public or private WhatsApp groups. A private group can only be joined through an invite, which is sent by the admin himself.

[3] https://en.wikipedia.org/wiki/Importance_of_religion_by_country.

Whereas for the public groups, WhatsApp allows users to enable joining these groups by means of an invite URL. This invite URL can be shared by any number of users and the users can join a group by clicking this URL. The invite URL has a specific structure of the form: chat.whatsapp.com/*. Many groups (e.g., for politics, sports, religion) share these invite URLs on social networks for easy visibility, hence increasing their community size and reach.

Group selection. Leveraging the unique structure of public WhatsApp groups, we used Facebook and Google search to find WhatsApp invite URLs. To ensure we find groups from Pakistani political parties, the URL was often used in conjunction with a political party's name or Pakistan. We also used Pakistani IP to ensure Google and Facebook give results from Pakistan as a priority. We also used the keyword "WhatsApp" along with political parties, names, or slogans. All of these queries have been logged for the convenience of new researchers in this domain and can be found online.[4]

Following these techniques, we found, in total 282 public Pakistani WhatsApp groups. A loose criterion was set to make sure the groups joined are in line with our research goals and will provide meaningful value to the overall dataset. We manually analyzed a group's bio, name, and profile pic. If any of these were not in line with our research goals, the group was removed from our dataset. Removed groups mostly were sales purchase groups, or groups made only for sharing jokes. Furthermore, we analyzed the remaining groups for activity over a week. In case a group did not have any significant activity, or the activity was not organic (e.g., a sales group), they were removed from our dataset. After these pruning steps, a total of 227 public groups were used to provide the analysis that follows.

WhatsApp data collection. To join these groups, we relied extensively on the tools created and provided by Garimella and Tyson (2018). These tools use selenium, enabling automated joining of public WhatsApp groups. WhatsApp messages are stored on a user's device in an encrypted form and is further protected by being placed in the root folder. The root folder is a secure folder and users do not have access to it. A rooted Android phone (having an unlocked bootloader and root folder) was used to obtain the decrypted database. The database is located in the data/data folder of the root drive. The media contents (images, videos) are stored online on WhatsApp servers and need to be accessed using the URLs provided in the database. Using these URLs, we downloaded images locally and decrypted them using a public tool.[5] This tool was not functional out of the box; hence, it was modified for our convenience and ease. During our data collection, it was observed that WhatsApp periodically deletes content from their servers. As a result, if a media file was not downloaded in a respectable amount of time, the media content could not be retrieved. Hence, a pipeline was created to extract data on weekly basis. An overview of our dataset can be seen in Table 1.

Table 1 Stats of our WhatsApp dataset, as collected from public Pakistani WhatsApp groups

Data type	Dataset
Groups	227
Admins	521
Users	18,475
All messages	23,895
Text messages	23, 895
Images	6699
Videos	1125
Audio	2741
Documents	1556

A detailed breakdown is given for our observed window between March 16 and April 9, 2020

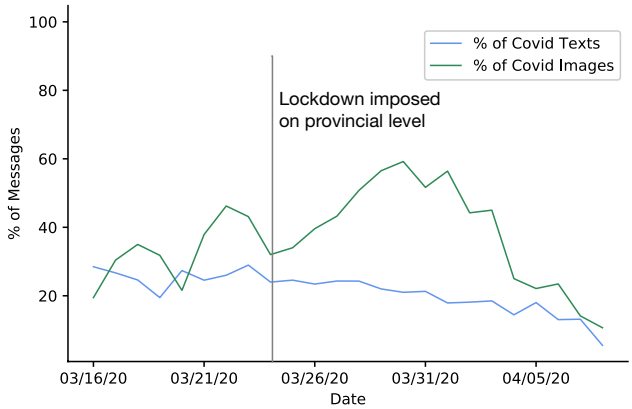


Fig. 1 Percentage of messages per day, for both text and images. A spike in images related to COVID-19 was seeing from 24/3 onward after lockdown was announced by Punjab on provincial level

3.2 Annotating COVID-19 text messages

COVID-19 text messages were extracted using a keyword-filtering approach. Rashed et al. (2020) provide a dictionary of keywords related to COVID-19. We used that and translated these words into equivalent Urdu terms. We also added small variations to these terms, like spelling mistakes, and multiple spellings to ensure that we capture a large dataset. Some of the sample keywords are: "corona," "covid-19," "covid," "covid19," and "coronavirus." It should be noted that, although this ensures that we get a large chunk of COVID-19-related messages, we still will miss some. This approach resulted in 5,039 text messages

[4] https://cutt.ly/8yXhxBd.

[5] https://github.com/ddz/whatsapp-media-decrypt.

Table 2 Distribution of images based on a group's political affiliation

Political affiliation	Num. of groups	Ambiguous (images) (%)	Information (images) (%)	Jokes (images) (%)	Misinfo. (images) (%)	Religious (images) (%)
Neutral	86	20	52	14	13	35
Opposition	103	19	41	6	10	22
Government	38	13	56	10	6	12

The affiliation is determined by a groups name, description, and profile picture

related to COVID-19 between March 16 and April 9, 2020. The dates roughly correspond with the first wave of COVID-19 in Pakistan.

Figure 1 gives a comparison of daily COVID-19-related and non-COVID-19-related messages in our dataset. One can observe the irregularity in the percentage of images related to COVID-19, compared to the stable flow of text messages. It was observed that rather than writing text messages, users found it more convenient to share images like news snippets and pictures from hospitals. In Fig. 1, the vertical dotted line represents lockdown being imposed on the provincial level. Before the lockdown was officially imposed, two spikes in percentage of images can be seen. These spikes relate to people sharing news about a mass spread event observed near the capital city. After the lockdown, an awareness campaign was started by the Government, in which images that contained helpful info related to COVID-19 were shared. Users in our groups actively shared these informative images. Another interesting debate that occurred during this time was the rulings on offering prayers in mosques. Since Pakistan is a Muslim majority country, many users shared sayings of scholars and news snippets related to this event as images. It is interesting to know that w.r.t. percentage — and in contrast to the trend seen for text messages — users were more inclined to share COVID-19-related images than non-COVID-19 images (see Fig. 1).

3.3 Annotating COVID-19 images

Twenty-five percent of the messages in our WhatsApp dataset contained images. We categorized images into COVID-19- and non-COVID-19-related images using manual annotators. Two annotators tagged a total of 6,699 images, between March 16 and April 9, 2020. This resulted in 4,490 non-COVID-19 images and 2,309 COVID-19 images. The two annotators had an inter-annotator agreement score of 98%. For the cases in which the annotators disagreed, the annotators were allowed to discuss the case and give a final label. Unlike text annotation, the annotators used a set of rules to identify the difference between COVID-19 images and non-COVID-19 images. If any of the following rules applied to an image, it was labeled as COVID-19:

1. Contained Coronavirus, COVID-19, or any other related terminology in Urdu or English.
2. Contained information relating to a lockdown or any restrictions being imposed/relaxed by the government on business or public/private institutions.
3. Contained 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.

3.4 Methodology for determining political affiliations

The groups joined are mainly from the political parties of Pakistan. Taking advantage of this, we look at COVID-19 as discussed by various parties in these groups. The groups are divided into three categories: 1) neutral, 2) opposition, and 3) government. To determine a group's affiliation, we look at the groups' name, profile image, and description. Many groups declare in their names the party to which they belonged to. The government and opposition groups are evident and comprise of any group which is affiliated with any government or opposition party. Whereas for neutral groups, these were constituency specific and did not openly declare affiliations toward any specific party. Out of our 227 groups, 86 were marked as neutral, 103 marked as affiliated to opposition parties, and 38 were marked as affiliated with the government.

To provide context, Table 2 shows the distribution of images shared per affiliation. We separate the images in several categories (as explained in Sect. 4). It should be noted that this does not reflect the overall government and opposition, as this is a skewed representation of the groups that are public in Pakistan. While keeping this caveat in mind, the government groups seem to share more information (rather than misinformation). In contrast, the opposition groups seem to share the largest amount of misinformation. The Neutral WhatsApp groups, on the other hand, share both information and misinformation messages in a large quantity. The Neutral WhatsApp groups also contain a large

Table 3 Precision, recall, and F1-score of logistic regression model as trained on 50,000 IMDB reviews

Sentiment	Precision	Recall	F1-score
Negative	0.874823	0.878203	0.876509
Positive	0.881691	0.878394	0.880039
Average/total	0.878313	0.878300	0.878303

amount of religious content (35%), which shows the importance of religion for the overall populace.

3.5 Methodology for understanding sentiment

Sentiment analysis is a useful tool for understanding user behavior. It also gives insights into the sentiment with which different type of messages (information, misinformation, etc.) are shared. Fake news often has a negative sentiment (Zaeem et al. 2020). To verify that this insight persists, we performed sentiment analysis on our dataset. Since our data came from political groups, it is also crucial to see the sentiment of the messages that also describes the general nature of the messages/discourse in political groups and whether these groups have a positive or negative sentiment about the COVID-19.

As there is a paucity of tooling available for the Urdu language, we train our own model to classify WhatsApp messages into positive and negative sentiment classes. As a training set, we rely on 50K IMDB reviews, which were translated into Urdu and released on Kaggle.[6] These reviews are divided into positive and negative classes. We use a Logistic Regression (LR) model provided by Scikit-learn (Pedregosa et al. 2011). The data are first transformed into a term frequency–inverse document frequency (TF-IDF) vector. The text, in this vectorized form, is fed to the classifier. After parameter optimization, we settled for the following parameters: the penalty used was "L2" with "saga" solver and balanced class weights. After training, we attain a precision and recall of 87.8% (see Table 3). Similarly, when doing inference, the text is first transformed into a TF-IDF vector and then sent to the LR model for inference.

We then select a threshold to classify a message as positive, negative, or neutral. Through manual experimentation, we decide the following ranges: A probability between 0.4 and 0.6 is considered neutral; under 0.4 is considered positive; and above 0.6 is considered negative.

4 RQ1: information sharing on COVID-19

To answer our first research question, we start by manually annotating our dataset into (overlapping) categories. The categories are explained in further detail below.

4.1 Message type categorization

We annotated text messages and images shared related to COVID-19 into five categories: *Information, Misinformation, Religious, Ambiguous, and Jokes/Satire*. To ascertain these categories, manual observation of the dataset was performed. It must be noted that a single message (text/image) can belong to multiple categories as the categories defined are not mutually exclusive. Each category is described below:

1. **Information:** This category consists of WhatsApp content that contains either of the two: News or COVID-19-related facts. Poynter's COVID-19 Facts database,[7] which compiles falsehoods detected by a large number of fact checking organizations, was used to determine the factual basis of the information shared. Furthermore, AFP Pakistan Fact Check[8] is used to verify news articles. The contents of the text or image, shared by WhatsApp users, are used to evaluate the falsehood or correctness of a message. Some claims were not present in the Poynter dataset. In such cases, if a piece of news is reported by a reputed news source, then it is labeled as "Information." A news source is considered reputed if it has a satellite news channel or newspaper at a national level. COVID-19-related facts are verified using WHO's COVID-19 Information and prevalent myths database.[9]
2. **Misinformation:** This is the inverse of the "Information" category. Similar to Information, Poynter COVID-19 Facts and Falsehoods database, AFP Pakistan, WHO's COVID-19 Information and COVID-19 Myth was used to verify a message. Any content, within a message, which is either verified to be misinformation or could not be verified as credible information is marked as misinformation
3. **Jokes/Satire:** A lot of users poked fun at the COVID-19 pandemic itself, or various COVID-19-related actions performed by local authorities using satire/ memes

   All the messages that contained this kind of information were marked as Jokes/Satire.

[6] https://www.kaggle.com/akkefa/imdb-dataset-of-50k-movie-translated-urdu-reviews/metadata.

[7] https://www.poynter.org/ifcn-covid-19-misinformation/.

[8] https://factcheck.afp.com/afp-pakistan.

[9] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters.

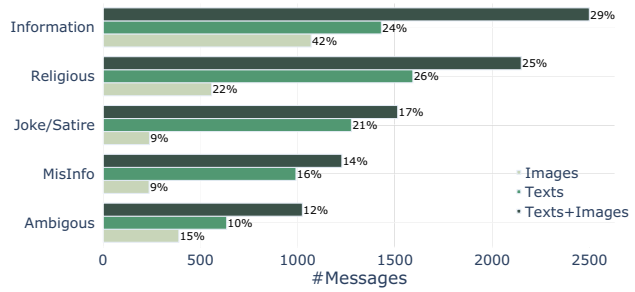


Fig. 2 Percentage of COVID-19 images (light green), texts (medium green), and texts + images (dark green) for each category. Notably, 14% of the total messages were labeled to be misinformation (color figure online)

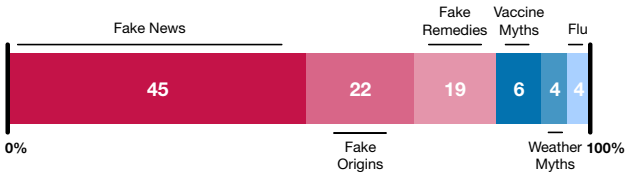


Fig. 3 The figure shows the breakdown of different types of misinformation as found in the dataset

4. **Religious:** A message is categorized as religious if it contains (i) references to religious texts, (ii) quotes of religious scholars (called *Maulana*, *Mufti*, or *Sheikh*), and (iii) emphasis on religious acts such as supplications, 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, the contact information of NGOs, donation requests, etc.

To maintain a consistent quality of annotations for images and text messages, the images were annotated by two annotators. Whereas the text annotations were first annotated by a single annotator, a randomly sampled 25% of the dataset was validated. Twenty-five percent of the sampled text messages were again annotated by another annotator. We found an 82% agreement score between the validating annotator and the original annotator, where a common label was counted as an agreement. It was found that the majority of the disagreement was between Information, Jokes/Satire, and Religious classes. This can be attributed to the mixed nature of texts, where jokes, religion, and information were often mixed in a single message. On the contrary, few disagreements were observed for messages that contained misinformation.

4.2 Message type analysis

A total of 2,309 images and 5,039 text messages were found that had COVID-19-related content between March 16 and April 9, 2020. Figure 2 shows the overall distribution of texts and images into COVID-19-related categories. The top category is that of information, having 29% of text messages and 24% of images. Religion is the second most popular category, with 26% of the text messages belonging to the religious category and 22% of Images. Pakistan is a country with a majority Muslim populace, which is reflected in the dataset. These statistics are reported without removing duplicates; we explore detecting and removing duplicates in sec 4.4. There is a variety of content related to religion, verses, and text from holy books (Quran), sayings of the Prophet (Hadith), sayings of religious scholars, and supplications. Furthermore, the impact of COVID-19 on religious life was also discussed vehemently, like the governments' decision to stop congregational prayers and rigorous testing of religious groups on proselytizing trips around the country.

The third most prevalent category is that of Joke/Satire. Out of all the text and image messages in our dataset, 17% contained Joke/Satire content. It was observed that users routinely ridiculed the government's actions and announcements. For example, a picture was seen circulating where an aircraft had a face mask on. Similarly, the governments' lack of initiative in closing borders was given a satirical spin and the increase in cases was blamed on officials. We also note the presence of a non-trivial amount of misinformation (14%). This means that 1/7 messages had misleading information related to COVID-19. Misinformation related to misleading news reporting or an image was shown out of context. For example, an image related to deaths in Italy was circulated, which in reality was an image from a film shoot. Overall, the percentage of misinformation is low but the types of misinformation were wide. We discuss different types of misinformation later in Sect. 4.3.

The "Ambiguous" category has the least amount of messages (12%). Many of the messages, in this category, contained requests for users to join their Facebook groups or subscribe to YouTube channels. These groups and channels were claimed to be related to COVID-19. Many requests for donations by different organizations and contact information of poverty-stricken COVID-19 patients were also shared. This was done to facilitate donations. Images depicting people with masks, or pictures of empty quarantine centers, and quarantined patients were shared. These images were labeled as "Ambiguous" as they did not contain any relevant information and were mostly out of context.

4.3 Misinformation

After classifying misinformation, the natural progression is to observe what type of misinformation was being shared by users. Misinformation was identified by relying on various fact checking organizations as mentioned in Sect. 3. Figure 3 shows the different types of misinformation present within the dataset. We detail their individual characteristics below.

Fake News. The most frequent form of COVID-19-related misinformation is in the form of fake news with 45% of misinformation texts. This includes fake news pertaining to COVID-19 positive tests and COVID-19-related deaths of world figures such as Ivanka Trump, Prince Williams, and even the 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 was 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 lake in Kazakhstan named "Corona," from which the virus came. A few Hollywood movies, namely "Contagion," "Resident Evil," and "I am Legend" along with the book "The Eye of Darkness" was 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-min breath-hold test to detect COVID-19, 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 economic leverage. Countries such as Israel, China, and USA were mentioned. This category makes up around 6% of the misinformation texts.

Weather Myths. Four percent of the misinformation items claim that the virus cannot survive in winter, summer, or rainy seasons, and that 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., USA), it did not have much salience in Pakistan.

4.4 Lifetime of COVID-19 messages

In this section, we explore the impact various types of COVID-19 messages have on the users of WhatsApp groups.

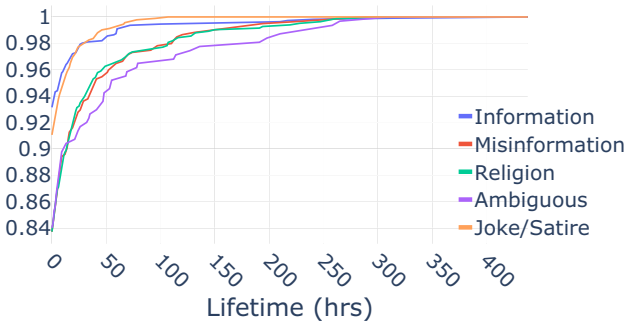


Fig. 4 Cumulative distribution functions (CDFs) of life of a COVID-19 Text message per category (note the broken *y*-axis: for better interpretability)

Since WhatsApp is a closed and private network, it is a difficult metric to calculate as WhatsApp does not contain any engagement metrics (unlike other platforms like Twitter that have retweet and like features). As a proxy, we therefore use the lifetime of a message within a given category. Lifetime is calculated using the first and last occurrence of a message (text or image).

To find the first and last occurrence of a particular text message, we perform identical String matching. To calculate the lifetime of an image, we group perceptually similar images using Facebook's PDQ hashing algorithm.[10] Two images were marked as similar if the Hamming distance between their hashes is below a threshold. We use a threshold of 70% as recommended by the PDQ authors. As a result, after finding similar images across the dataset, the images' first occurrence and last occurrence were recorded.

The lifetimes of text messages are shown in Fig. 4, whereas the lifetimes of both text and images are presented in Table 5. Every category can be identified by its unique variance and mean measure. The jokes/satire category is the most short-lived. This can be attributed to the fact that the category has the largest unique set of content, hence low repetition. Furthermore, jokes/satire content is generally dictated by events and usually dies out quickly as new events happen. An alarming observation is the lifetime of misinformation. Misinformation has the highest lifetime for both text (7 h) and images (5 h). Since WhatsApp is a closed network, there is limited professional moderation, which makes the long-lived life of misinformation messages quite alarming. Our findings are in line with observations made by other researchers, specifically (Vosoughi et al. 2018; Garimella and Eckles 2020), showing that misinformation tends to live longer and penetrate deeper than the rest.

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

4.5 Lifetime of misinformation

Due to its seriousness, we further inspect the lifetime of misinformation messages. Misinformation was divided into subcategories as mentioned in Sect. 4.3. To understand the temporal properties of various types of misinformation, we calculate the life of all the messages that contained different types of misinformation. The breakdown of temporal properties can be seen in Table 5. Here, we observe that weather myths are the longest lived, and have a staggering 26hrs of lifetime, whereas fake news has a short lifetime, roughly 4hrs. News shares the same properties as Jokes/Satire: they are short-lived and tend to be replaced by other news quickly. Hence, it has a shorter lifetime than other categories of misinformation. Fake remedies and weather myths have the highest lifetimes. This can be attributed to the nature of these categories as they are not time-bound and tend to stay constant over time. Hence, the type of misinformation is important in the impact it has and the time it stays relevant.

5 RQ2: user behavior and political inclination

5.1 Political affinity and misinformation trends

A WhatsApp group's political affiliation was deriving as described in Sect. 3.4. As a result, each group is labeled as neutral, belonging to opposition parties, or affiliated with the government. This allows us to understand the political side of COVID-19 messages and whether having a specific political inclination affects the type of messages a user is bound to share.

5.1.1 COVID-19 messages

Concerning political affiliation and COVID-19 messages, we observe that neutral groups have the highest percentage of misinformation messages shared (Fig. 5). The largest amount of correct health information was shared by groups affiliated with the government. This can be attributed to government efforts to share information relating to the pandemic. As they are government supporters, they are more likely to share government vetted information, resulting in a large amount of health information. Unlike neutral and opposition groups, the government groups are seen to be posting the least amount of religious messages related to COVID-19. The number of jokes about COVID-19 is equally spread across the groups. This perhaps highlights the need for governments to come up with better ways to promote health information in non-government leaning groups.

5.1.2 Misinformation

In Sect. 4.3, we looked at the overall distribution of textual misinformation across all the groups. Textual misinformation was categorized into fake news, fake origins, fake remedies, vaccine myths, weather myths, and flu. We next explore what differences exist in this distribution given a group's political inclination. Figure 6 shows the distribution of misinformation messages within groups having a specific political affiliation. This shows roughly similar trends across the different political affiliations. The majority of the messages are related to fake news, closely followed by fake origins and fake remedies. While comparison with flu is the least discussed category. This trend is the same irrespective of the political inclination.

5.2 Individual user's behavior

A WhatsApp groups user base can be split into two categories: 1) *producers*, and 2) *consumers*. Producers are in the minority and are the most vocal in sharing messages and producing new content. On the other hand, consumers silently consume messages, rarely interacting. The producers and consumers can be observed in the dataset overview provided above (Table 1). This section is dedicated to analyzing the user behavior and understanding if there is any deliberate spread of misinformation (disinformation).

To profile user activity, we take advantage of "UpSet" plots.[11] UpSet plots give a clean and easy-to-read view of the set overlaps within a dataset. In our context, every set is a unique user and the contents are the type of messages being shared by that user. The set overlap represents the similarity of content being shared by users. As seen in Figs. 7 and 8, the bottom matrix (known as combination matrix) represents the intersection of the sets (users) across various COVID-19 classes. The bars on top show the number of sets (users) within a given intersection. To see the number of users posting content related to a single COVID-19 category, the left bar should be observed.

Figure 7 represents the text messages shared by individual users in various groups. By looking at the combination matrix, we observe the general user content sharing trends. For example, the most exclusively shared category is "Ambiguous." This shows that most users like to share neutral content related to COVID-19, having no informational value. These are mostly, calls of concern or donation efforts. This can be equally attributed to the excitement and concern of the users. The second and third categories exclusively share religious and misinformation content. This deviates

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

Table 4 Lifetimes of COVID-19-related texts and images shared on WhatsApp

Label	Num. texts	Mean (h)	Std dev (h)
Text messages			
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
Label	Num. images	Mean (h)	Std dev (h)
Images			
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

Misinformation tends to have the highest mean lifetime

Table 5 Lifetime of misinformation texts shared on WhatsApp

Label	Num. texts	Mean (h)	Std dev (h)
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

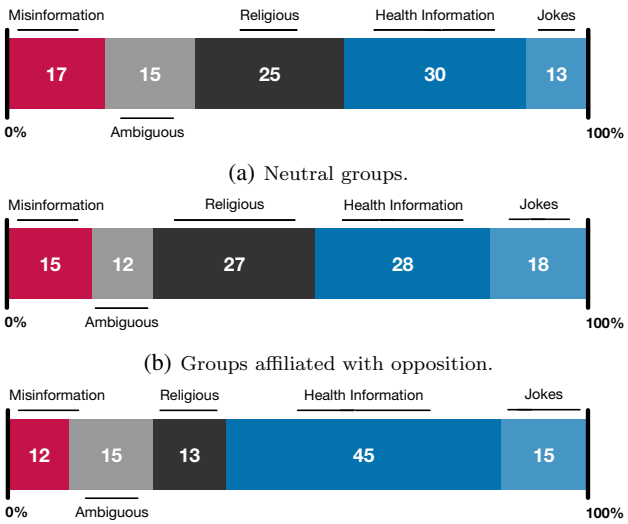


Fig. 5 Percentage of COVID-19-related text messages and images in different groups. The groups are differentiated in sub-graphs on the basis of their political inclinations

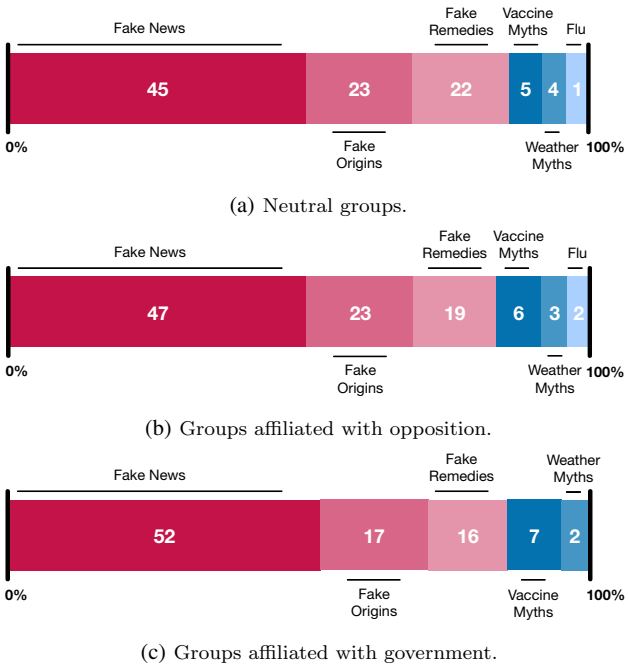


Fig. 6 The graphs show the breakdown of the percentage of COVID-19 misinformation per category shared in groups with different political orientations. The sub-graphs represent the distribution of messages based on a group's political affiliation in which the message was shared. The political affiliation is extracted based on a group's name, image, and description

largely from what we have seen in Image sharing trends. For text messages, users prefer to share misinformation only, whereas, for images, we do not see any such trends.

Figure 8 depicts the user behavior when sharing images related to COVID-19. Unlike text messages, the majority

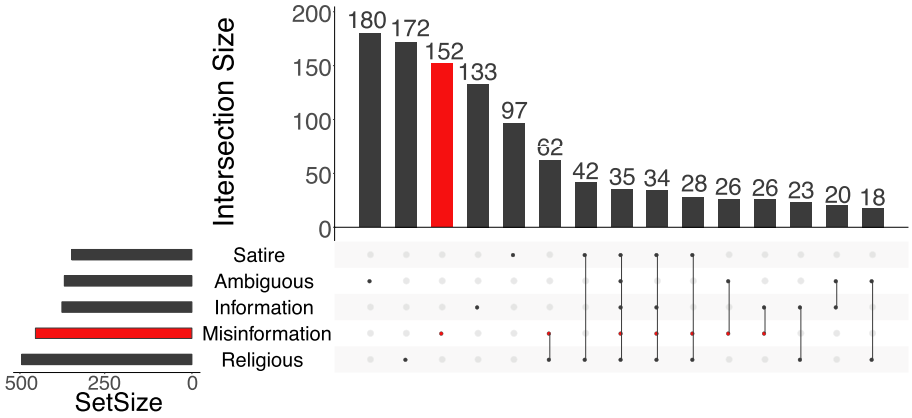


Fig. 7 UpSet plot of top 15 intersection sets for users posting COVID-19-related text messages in public WhatsApp groups. More users appear to share texts that belong to a single category

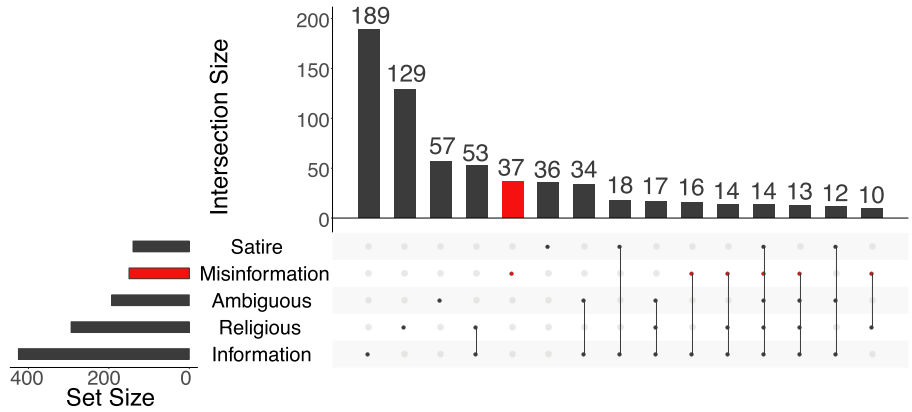


Fig. 8 UpSet plot for users posting COVID-19-related images. Only the top 15 intersection sets are visualized. A lot of users are sharing information and religious content, whereas some share misinformation

of the images being shared by users contain facts and correct information, whereas very few users are sharing misinformation. Only 37 users exclusively shared misinformation. On the contrary, 189 users shared images that contained fact. If we look at mixed content sharing, we observe that 67 users shared a mix of content, sharing misinformation along with some other type of content.

We next try to understand if there is any organized spread of misinformation, or if a single instance of misinformation is being shared more widely than the rest. Note, we have already calculated the similarity between different images PDQ. Based on this similarity, we find that only 23 images were shared more than once, out of which 8 images were shared more than 5 times. These 8 images may be a pointer toward either a widespread and common misinformation or a disinformation campaign.

6 RQ3: cross network information spread

We conjecture that there may be flows of information across WhatsApp and other social networks. Thus, we next explore if information passes between Twitter and WhatsApp. This means we can further rely on new metrics provided via Twitter, such as likes and retweets.

6.1 Methodology

Since WhatsApp is a closed network, we perform a cross network study of data flow. Twitter is an ideal choice for this type of study as, unlike WhatsApp, it is a major conduit of information related to user interactions. We obtain around 0.8 million tweets relate to COVID-19 for Pakistan, using relevant hashtags. Examples of some relevant hashtags are CovidPakistan, CoronaPak, and CoronaPakistan. Note that this is only a subset of activity and should not be generalized for the whole population. In order to understand the flow of information and obtain user interactions for the content in our dataset, we map images between March 16 and April 9, 2020 for both WhatsApp and Twitter. We then download all the images from both WhatsApp and Twitter and use PDQ hashing and hamming distance to find similar images.

6.2 Cross platform image spread

A total of 67,119 images were downloaded from our Twitter dataset. After downloading the images, we generate PDQ hashes for each one. We then cluster together similar images. We then use Hamming distance to find similarity between the PDQ hashes generated for WhatsApp images and Twitter images. Around 1500 were found to be similar between the

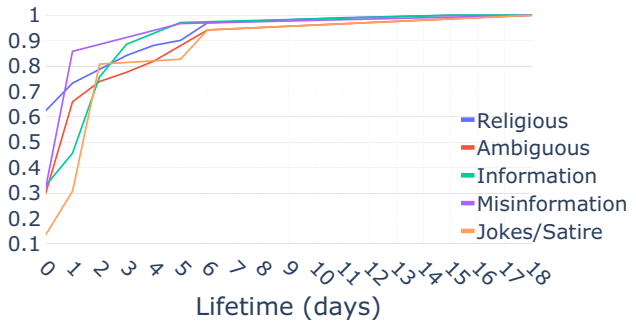


Fig. 9 CDFs of life of an image, along with content type, as seen on Twitter. Twitter tends to hold a message alive for a couple of days. A healthy trend is that images in the "Information" category live the longest on Twitter

two datasets. Table 6 gives a breakdown of the images that were found to be similar.

After finding similar images, we calculate the lifetime of a message on Twitter. This is the temporal distance between the time an image was first observed on Twitter and the last interaction on it. Interactions include a like, a tweet containing that image, a reply or a retweet. We find that the highest number of retweets is for misinformation. This can be attributed to the nature of misinformation, as it is unique and insightful and users tend to retweet it faster than other categories.

Although misinformation tends to be retweeted the most, it still has the lowest lifetime. This can be contrasted to WhatsApp where the lifetime of misinformation is the longest. This can be the result of the continuous interaction of users with a tweet. For a tweet containing misinformation, users could be negating the information and hence, reducing its life compared to others. Figure 9 shows the CDF for the life of a message on Twitter. Most of the categories have a long tail. This long lifetime is inverse of what was observed on WhatsApp. This can be attributed to the nature of WhatsApp, where old messages are replaced by new (messages are displayed in chronological order) and conversation chains are few. This nature of WhatsApp results in messages being alive for no more than a few hours (Table 4). In contrast, Twitter posts can be kept alive for days (Table 6). This is because Twitter offers customized feeds to users that deviate from chronological order. This difference in the nature of the two social networks is the reason behind the stark difference in the lifetime of a message on both platforms.

Figure 10 depicts the flow of information between WhatsApp and Twitter. The figure also gives 3 examples of how images originated from WhatsApp and were later to become widely tweeted. Most of the images found common between WhatsApp and Twitter, first occurred on WhatsApp rather than Twitter. This may be specific to our dataset and should not be taken as a generalized trend. On average, an image that has appeared on WhatsApp tends to appear 4 days later

Table 6 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

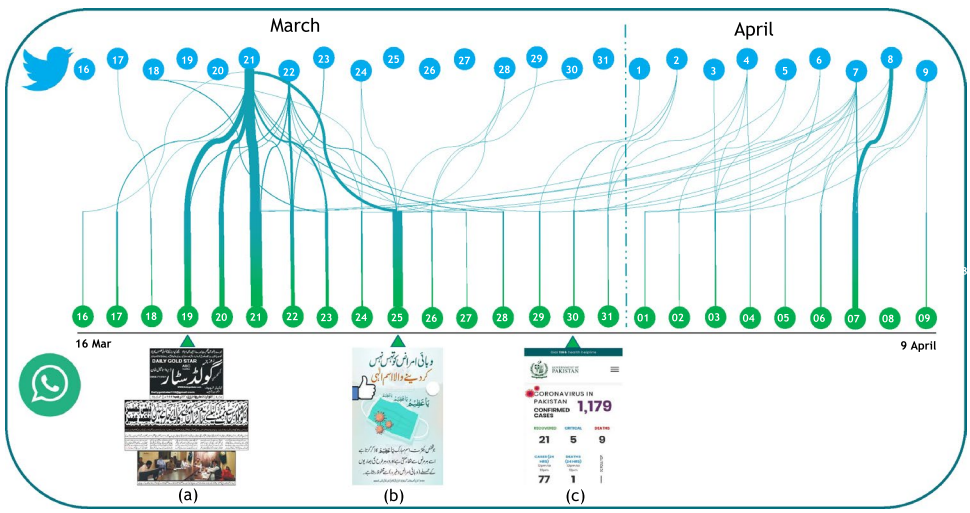


Fig. 10 COVID-19 images' temporal flow across WhatsApp and Twitter (a line's thickness depicts the number of images flowing across). *Some Observations*: **a** a news snippet originates from WhatsApp on March 19th and is seen on Twitter on 21st; **b** religious supplication to fight COVID-19 is observed on WhatsApp 2 days earlier than on Twitter; **c** official stats of COVID-19 patients seen on March 30th on WhatsApp earlier than on Twitter

on Twitter. This shows the importance that WhatsApp holds in content creation and directing the direction of discussion on online social networks for the Pakistani populace. These findings are important, in the context of the results observed in Table 6, where the majority (30%) of the messages that are common between the two platforms are misinformation, compared to 12% being information.

6.3 Bots on WhatsApp

Bot activity is actively increasing across social networks (Shao et al. 2018). They are even sometimes deployed to drive the narrative in the political Twitter-sphere. Detecting WhatsApp bots is not a mature field, and there have not been any proposed methods of finding them as of yet. In contrast, a lot of work has been done on bot detection for Twitter (Gilani et al. 2019, 2017). In Sect. 6, we analyzed the flow of information between WhatsApp and Twitter. As a result, we created a mapping of which content in WhatsApp was shared by which users on Twitter. Leveraging this information, we try to understand what type of content is being shared by WhatsApp users, is also being shared by bots on Twitter.

Using the Yang et al. (2020) bot detection methodology, we perform bot classification of the Twitter users that are mapped with WhatsApp data. The authors have provided a trained model under the name of Botornot-v4. The model was trained on a total of 94,124 bots and 43,396 human accounts. An advantage of using this dataset is that it does not take into account the content of a tweet, but rather looks at an account's metadata and also uses derived features from that account's metadata. The result is that irrespective of the language in a given tweet, we can ascertain with a reasonable probability if an account is a bot. These features are then submitted to a random forest for prediction. We used the service using a Python library provided by the authors.[12]

Out of the 470 accounts on Twitter, a total of 33 accounts were classified as bots. These accounts mostly share correct information or religious content. This suggests that these are social bots but are not interested in the spread of misinformation. Instead, the amount of misinformation shared by these accounts over the period of 3 weeks can be attributed to the infodemic. Out of all the messages being propagated by these bots, 14.8% were classified as misinformation. This is in-line with the overall trend observed by us in WhatsApp data, further enhancing our assumption that these bots are not sharing misinformation on purpose. That said, it must be noted that this does not represent the full scope of Twitter or WhatsApp activity, and this could be the case for our specific dataset only.

7 RQ4: sentiment around COVID-19 messages

We next ask what type of sentiment is exhibited when a user is sharing a message about COVID-19. More specifically, we look at the text messages classified earlier and try to observe the sentiments in each category of message. The categories are defined in the methodology section above.

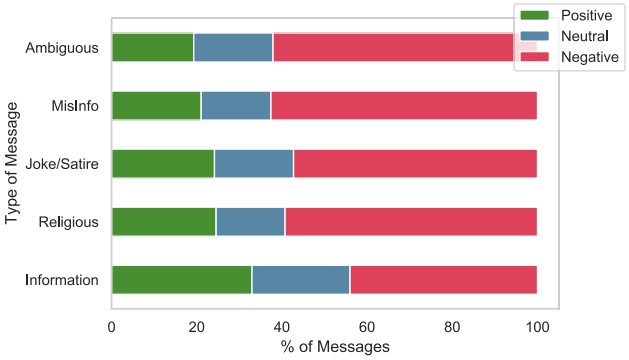


Fig. 11 Distribution of sentiment across different message types. The prevailing sentiment is negative

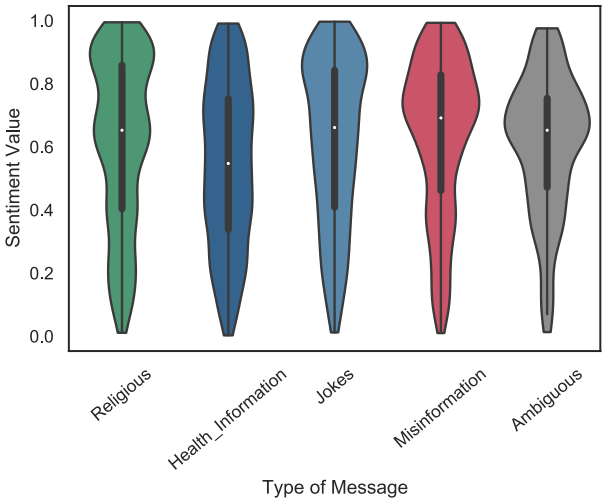
7.1 Sentiment across message category

Figure 11 presents the sentiment breakdown for each category of message. Interestingly, most messages have a negative sentiments. Relative to other categories, information messages have the highest fraction of positive and neutral sentiment. This can perhaps be attributed to the nature of the information category. As these messages mostly consist of news articles, headlines, or factual posts. As a result, this type of content tends to be neutral. For example, a headline containing information regarding the number of COVID-19 cases on a specific date gets classified as neutral (because it is only mentioning a fact and does not have any negative or positive sentiment). As a caveat, it should be noted that the algorithm used for classification does not understand COVID-19 terminology.

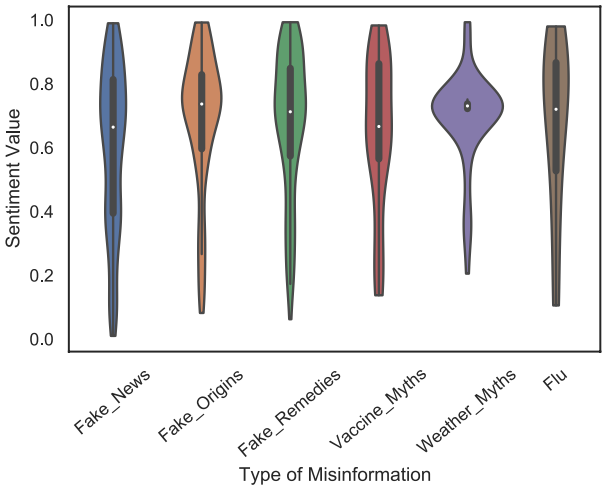
To complement this above, Fig. 12 presents the full distribution of sentiment scores as a violin plot. We observe that health information is spread across the spectrum equally and the interquartile range is almost in the center. Whereas misinformation seems to have the most negative sentiment, with the kernel density estimation (KDE), plotting the highest density above 0.6.

[12] https://github.com/IUNetSci/botometer-python.

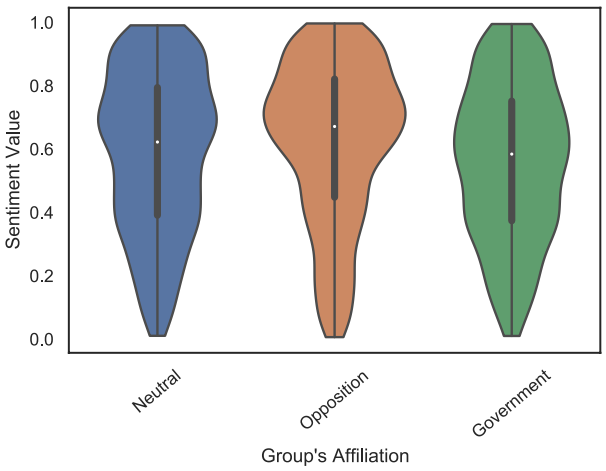
Fig. 12 Violin plots showing results of text sentiment analysis for all COVID-19 messages. Within the violin plot, box plots are rendered. The sentiment analyzer gives values between 0 and 1. A sentiment value greater than 0.6 is considered negative, and a sentiment value lower than 0.4 is considered positive, whereas the values in-between are considered neutral



(a) Sentiment values for all COVID-19 messages categorized into 5 classes.



(b) Sentiment values for all COVID-19 misinformation messages.



(c) Sentiment values, on the basis of a groups political affiliation.

7.2 Sentiment across misinformation

We further group the text-based misinformation into categories, as explained in Sect. 4.3. Next, we try to understand what is the general sentiment of the messages that contain misinformation. As seen in Fig. 12b, almost all the messages are classified as negative. This is in line with the overall message trend. The highest level of negative sentiment is seen in the weather myths category. Weather myths are a declaration that COVID-19 will be gone when the weather turns warmer. Weather myths are closer to being a conspiracy theory, where people negate the seriousness of COVID-19 and propose that with warmer weathers COVID will end. They therefore mostly talk negatively about the authorities for not understanding this, or about people they consider are profiting from COVID-19.

Interestingly, fake news is more often associated with a positive sentiment. This may be related to the way the classifier understands the text messages. For instance, a fake news piece was shared telling people that "A famous footballer was found to have corona" (translated from Urdu). This message was classified as positive, rather than negative as "famous footballer" is a positive word while "corona" is not detected as negative. This explains why most of the fake news is classified as positive. Furthermore, note that news is often told subjectively, and with the classifier unable to understand if "corona" is negative or positive, it does not attribute negative nature to fake news.

7.3 Sentiment of political statements

Political polarization is rampant (Jiang et al. 2020). Pakistan has many political parties. Thus, rather than dividing the populace into right or left wing, we divide groups into affiliated with government and opposition groups as explained in Sect. 3.4. To explore the sentiment of these two opposing groups, Fig. 12c displays violin plots of the sentiment scores for groups that fall into each political affiliation. Interestingly, the groups associated with opposition parties have a negative proclivity as compared to the rest. Groups associated with the government tend to have positive or neutral conversations around COVID-19. This can be associated with the opposition's tendency to critique government policy regarding COVID-19 (in contrast

to government supporters who praise current policies). As seen in Fig. 6b, opposition groups in our dataset tend to share more misinformation than government groups too.

8 Conclusion

Understanding information spread about the pandemic on social media enables us to tap into the pulse of modern societies. In this study, we have analyzed the spread of information through WhatsApp messages in the context of Pakistan political groups. To the best of our knowledge, this is the first study that provides insights into how politics, the infodemic, and misinformation play a role in Pakistani society during the COVID-19 pandemic. Our work has made a number of key findings. We found that around 14% of messages are misinformation and that political party affiliated groups do play a role in the dissemination of misinformation. In the context of the pandemic, it was found that opposition parties tend to share less information and more misinformation, whereas the opposite was observed for leading parties. We also identified overlap between WhatsApp and Twitter and found that information originates earlier on WhatsApp as compared to Twitter. Prior to common belief, we further observed that bots on Twitter are not excessively involved in spreading misinformation. We, of course, emphasize that our results are based on a small subset of the overall conversations taking place on WhatsApp. Thus, as part of our future work, we wish to scale-up our measurements to explore if our observations generalize. This is particularly important in confirming the impact that political affiliation has on behavior.

Acknowledgements This work is supported by the UK EPSRC, under Grant EP/S033564/1 and facebook (Grant No. FY20PP013).

References

Badawy A, Addawood A, Lerman K, Ferrara E (2019) Characterizing the 2016 Russian IRA influence campaign. Soc Netw Anal Min 9(1):1–11

Bhatnagar S, Choubey N (2021) Making sense of tweets using sentiment analysis on closely related topics

Boadle A (2018) Facebook's Whatsapp flooded with fake news in brazil election

Bovet A, Makse HA (2019) Influence of fake news in twitter during the 2016 us presidential election. Nat Commun 10(1):1–14

Chen E, Lerman K, Ferrara E (2020) Tracking social media discourse about the Covid-19 pandemic: development of a public coronavirus twitter data set. JMIR Public Health Surveill 6(2):e19273

Cinelli M, Conti M, Finos L, Grisolia F, Novak PK, Peruzzi A, Tesconi M, Zollo F, Quattrociocchi W (2019) (Mis) information operations: an integrated perspective. arXiv preprint arXiv:1912.10795

Cinelli M, Quattrociocchi W, Galeazzi A, Valensise CM, Brugnoli E, Schmidt AL, Zola P, Zollo F, Scala A (2020) The COVID-19 social media infodemic. arXiv preprint arXiv:2003.05004

David MJL, Matthew AB, Yochai B, Adam JB, Kelly MG, Filippo M, Miriam JM, Brendan N, Gordon P, David R et al (2018) The science of fake news. Science 359(6380):1094–1096

de Freitas Melo P, Vieira CC, Garimella K, de Melo POSV, Benevenuto F (2019) Can WhatsApp counter misinformation by limiting message forwarding?

Evangelista R, Bruno F (2019) Whatsapp and political instability in brazil: targeted messages and political radicalisation. Internet Policy Rev 8(4):1–23

Garimella K, Eckles D (2020) Images and misinformation in political groups: evidence from WhatsApp in India. arXiv preprint arXiv:2005.09784

Garimella K, Tyson G (2018) Whatsapp, doc? A first look at Whatsapp public group data. In: Twelfth international AAAI conference on Web and Social Media

Gilani Z, Farahbakhsh R, Tyson G, Wang L, Crowcroft J (2017) Of bots and humans (on Twitter). In: Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017, pp 349–354

Gilani Z, Farahbakhsh R, Tyson G, Crowcroft J (2019) A large-scale behavioural analysis of bots and humans on twitter. ACM Trans Web 13(1):1–23

Goel V (2018) In India, Facebook's WhatsApp plays central role in elections

Grinberg N, Joseph K, Friedland L, Swire-Thompson B, Lazer D (2019) Fake news on twitter during the 2016 US presidential election. Science 363(6425):374–378

Habib A, Asghar MZ, Khan A, Habib A, Khan A (2019) False information detection in online content and its role in decision making: a systematic literature review. Soc Netw Anal Min 9(1):1–20

Iosifidis P, Nicoli N (2020) The battle to end fake news: a qualitative content analysis of Facebook announcements on how it combats disinformation. Int Commun Gazette 82(1):60–81

Javed RT, Shuja ME, Usama M, Qadir J, Iqbal W, Tyson G, Castro I, Garimella K (2020) A first look at Covid-19 messages on Whatsapp in Pakistan. arXiv e-prints, pages arXiv-2011

Jiang J, Chen E, Yan S, Lerman K, Ferrara E (2020) Political polarization drives online conversations about Covid-19 in the united states. Hum Behav Emerg Technol 2(3):200–211

Kouzy R, Jaoude JA, Kraitem A, El Alam MB, Karam B, Adib E, Zarka J, Traboulsi C, Akl EW, Baddour K (2020) Coronavirus goes viral: quantifying the Covid-19 misinformation epidemic on Twitter. Cureus 12(3):66

Lokniti CSDS (2018) How widespread is WhatsApp's usage in India?

Maros A, Almeida J, Benevenuto F, Vasconcelos M (2020) Analyzing the use of audio messages in Whatsapp groups

Newman N, Fletcher R, Kalogeropoulos A, Nielsen RK (2019) Reuters Institute Digital News Report 2019 . Reuters Institute for the Study of Journalism

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830

Perrigo B (2019) How volunteers for India's ruling party are using WhatsApp to fuel fake news ahead of elections

Purnell N (2020) Facebook's whatsapp battles coronavirus misinformation

Rashed SK, Frid J, Aits S (2020) English dictionaries, gold and silver standard corpora for biomedical natural language processing related to sars-Cov-2 and Covid-19

Resende G, Messias J, Silva M, Almeida J, Vasconcelos M, Benevenuto F (2018) A system for monitoring public political groups in Whatsapp. In: Proceedings of the 24th Brazilian symposium on multimedia and the Web, pp 387–390

Resende G, Melo P, Reis JCS, Vasconcelos M, Almeida JM, Benevenuto F (2019a) Analyzing textual (mis)information shared in

WhatsApp groups. In: Proceedings of the 10th ACM conference on web science

Resende G, Melo P, Sousa H, Messias J, Vasconcelos M, Almeida J, Benevenuto F (2019b) (Mis) information dissemination in Whatsapp: gathering, analyzing and countermeasures. In: The World Wide Web conference, pp 818–828

Saha P, Mathew B, Garimella K, Mukherjee A (2021) "Short is the road that leads from fear to hate": fear speech in Indian Whatsapp groups. arXiv preprint arXiv:2102.03870

Shao C, Ciampaglia GL, Varol O, Yang K-C, Flammini A, Menczer F (2018) The spread of low-credibility content by social bots. Nat Commun 9(1):1–9

Sharma K, Qian F, Jiang H, Ruchansky N, Zhang M, Liu Y (2019) Combating fake news: a survey on identification and mitigation techniques. ACM Trans Intell Syst Technol 10(3):1–42

Sharma K, Seo S, Meng C, Rambhatla S, Dua A, Liu Y (2020) Coronavirus on social media: Analyzing misinformation in Twitter conversations. arXiv preprint arXiv:2003.12309

Shu K, Sliva A, Wang S, Tang J, Liu H (2017) Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor Newsl 19(1):22–36

Shu K, Wang S, Lee D, Liu H (2020) Mining disinformation and fake news: concepts, methods, and recent advancements. arXiv preprint arXiv:2001.00623

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

Starbird K, Maddock J, Orand M, Achterman P, Mason RM (2014) Rumors, false flags, and digital vigilantes: misinformation on twitter after the 2013 Boston marathon bombing. In: IConference 2014 proceedings

Trevisan M, Vassio L, Drago I, Mellia M, Murai F, Figueiredo F, da Silva APCo, Almeida JM (2019) Towards understanding political interactions on Instagram. In: Proceedings of the 30th ACM conference on hypertext and social media, pp 247–251

Two Billion Users (2020) Two billion users—connecting the world privately. https://blog.whatsapp.com/two-billion-users-connecting-the-world-privately

UK says WhatsApp (2017) UK says WhatsApp lets paedophiles and gangsters operate beyond the law. Reported: 3 (2017). https://www.reuters.com/article/us-britain-security-whatsapp-idUSKCN1C8165

Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359(6380):1146–1151

Yadav A, Garg A, Aglawe A, Agarwal A, Srivastava V (2020) Understanding the political inclination of Whatsapp chats. In: Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, pp 361–362

Yang K-C, Varol O, Hui P-M, Menczer F (2020) Scalable and generalizable social bot detection through data selection. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 1096–1103

Zaeem RN, Li C, Barber KS (2020) On sentiment of online fake news. In: 2020 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). IEEE, pp 760–767

Zarei K, Farahbakhsh R, Crespi N, Tyson G (2020) A first Instagram dataset on Covid-19. arXiv preprint arXiv:2004.12226

Zhang X, Ghorbani AA (2020) An overview of online fake news: characterization, detection, and discussion. Inf Process Manag 57(2):102025

Zhou X, Zafarani R (2018) Fake news: a survey of research, detection methods, and opportunities. arXiv preprint arXiv:1812.00315

Zollo F, Quattrociocchi W (2018) Misinformation spreading on Facebook. In: Complex spreading phenomena in social systems. Springer, pp 177–196