(Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures

Ankit Agrawal

Seminar on Data-driven Approaches on Understanding Disinformation

Lecturer: Savvas Zannettou, Yang Zhang

July 1, 2020

Contents

1 Abstract

In Brazil, 48% of the population uses WhatsApp to share and discuss news rising a serious concern if it can be host to groups interested in disseminating misinformation, especially as part of articulated political campaigns. In this work, the authors have analyzed information dissemination and misinformation footprint on publicly available WhatsApp groups related to two major political events held in Brazil i.e. national truck drivers' strike and the Brazilian presidential campaign.

2 Introduction

WhatsApp is the most popular messaging application in the world having more than 1.5 billion active users (Constine (2018)). Given its popularity and the usages, it has also become one of the major hosts for the dissemination of misinformation, especially for political campaigns. In this paper, authors focus on publicly accessible groups i.e. groups for which invitation link has been activated focusing on political campaigns. The paper provides a large-scale investigation of information dissemination in these WhatsApp groups. Here the data has been gathered from the content shared in these publicly available groups for the periods corresponding to two major social mobilization events that took place in Brazil: (i) a national truck drivers' strike (May 21st to June 2nd, 2018); and (ii) the first round of the 2018 Brazilian general elections campaign (August 16th to October 7th, 2018), with 141 and 364 groups monitored, respectively. In the analysis provided by the authors, we first see the details of the contents that have been shared in these groups, user interactions, and the dissemination of this information. And further, we can see the details about the misinformation present in these contents especially for the images shared and its propagation to and from the Web.
In the next section, we discuss how the data was acquired followed by characterizing the contents of this data in section 4. Then, the analysis on the dissemination of misinformation can be seen in section 5. Finally, Section 6 concludes the paper findings.

3 Data Acquisition

As we know WhatsApp is end-to-end encrypted, which means apart from the people involved, no one can access the messages. However, if for a group the invitation link has been activated and is shared on the web makes it possible for anyone to access or join the group, hence making it publicly accessible. To identify these publicly accessible groups, the authors used the part of the invitation link URL i.e. "chat.whatsapp.com" as a search query to Web and other social media platforms. They furthered filtered it using the keywords from the 2018 Brazilian election dictionary. A total of 1828 valid invitation links were gathered after this process.
Each group was then joined using cell phones and a tool developed by (Garimella & Tyson (2018)). For each group all the data shared was then stored in a database. Table 1 provides an overview of the data content gathered and the number of messages per categories (text, image, video, and audios). Here we

	Truck Drivers' Strike	Election Campaign
#Groups	141	364
#Total Users	5,272	18,725
#Total Messages	121,781	789,914
#Text Messages	95,424	591,162
#Images	11,610	110,954
#Videos	9,752	73,310
#Audios	4,995	14,488
#URLs	11,728	92,654

Table 1: Overview of the Dataset

see that image is one of the most frequent types of content in both datasets, reaching roughly 10% and 15% of all content shared. Also, the initial analysis done by a set of journalists in these data suggested that images are the main source of misinformation among all data types. In the later section, we will use images to study the dissemination patterns.

4 Message content and dissemination

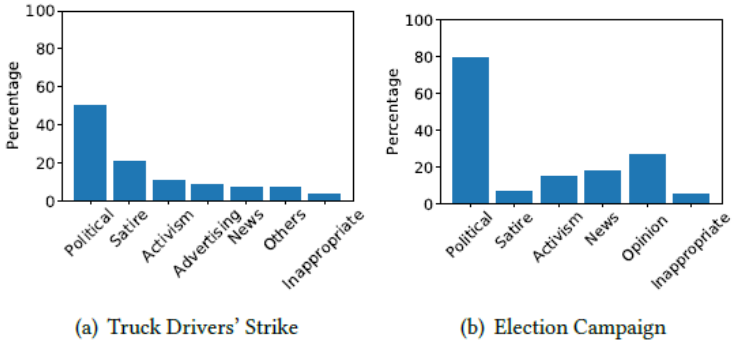
In this section, we see the distribution of image categories, and how it is propagated to and from the Web.

4.1 Characterizing WhatsApp images

WhatsApp groups are a general space for discussion for various topics ranging from politics, science to advertisements. To understand the different kinds of images spread, we first need to categorize our monitored images using content labeling. In this section, we also see how these images are spread across the web and how WhatsApp acts similar to other big social media networks like Twitter, Facebook, etc.

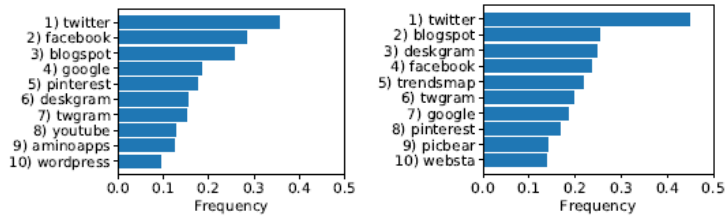
4.1.1 WhatsApp images content labeling

To categorize the images based on content, three volunteers were asked to label a sample of most shared images during each event. For truck drivers' strike period, a selection of top-20 most shared images on each day, with a total of 220 images, similarly for the election campaign period, a selection of top-100 most shared images per day was considered. Duplicates images were removed using the Perceptual Hashing (pHash) algorithm (Monga & Evans (2006)). Volunteers are giving a taxonomy guideline document with instruction were able to classify images into these categories (political, news, advertising, satire, activism, opinion, others). Figure 1 shows the distributions of image categories for both events, we can see that most images are related to politics i.e. 50% during the truck drivers strike and 80% during the election campaign.



(a) Truck Drivers' Strike (b) Election Campaign

Figure 1: Distributions of image categories.



(a) Truck Drivers' Strike (b) Election Campaign

Figure 2: Most popular domains for images shared on WhatsApp publicly accessible groups

4.1.2 WhatsApp Images on other Websites

Here the authors try to find the extent to which images shared in these groups have also appeared on other websites and social media platforms. Here it uses Google Image Search to lookup for the presence in other websites for the same images found in our monitored groups. Figure 2 shows the popular domains found via a Google Image search for the images present in the event duration. We can also see that social media platforms like Twitter, Facebook, Aminoapps, and Pinterest are among the most frequent domains where these monitored images were also posted where Twitter being the leader for both the events.

4.1.3 Network Structure

The authors have performed an analysis to get the relations between these monitored groups and the users related to these groups. We see that few users act as a bridge between multiple highly active groups allowing the easy spread of the images. The analysis makes a strong claim on how WhatsApp acts similar to many other social networks such as Twitter or Facebook and connecting thousands of users, it has the potential to make a piece of information viral.

4.2 Propagation Dynamics

In this section, the authors have discussed the propagation of these monitored images in the WhatsApp and to and from the outer web by the means of two metrics, namely lifetime and burst time. The lifetime of an image is the time interval between the first and last instance when an image was shared, whereas burst time is the time interval between consecutive shares of the same image. We can see from the analysis that: lifetimes for 20% of the images were under 1 hour, for 40% of them it was under 20 hours whereas for 30% of them it was found to be exceeding 100 hours. Similarly, for burst time it was found that 40% of the images were reshared within up to 120 and 100 minutes during the strike and election events respectively. They also analyzed for the first appearance of these images, and it was found that 14% of the images were shared on the web and WhatsApp on the same day whereas 6% of them were shared on WhatsApp first.

5 Misinformaiton on WhatsApp

In this section, we look at the amount of misinformation present in these images shared on the monitored groups. We also see how we can identify the misinformation manually as well as automatically. And we see the characteristics and propagation dynamics for this misinformation identified images.

5.1 Misinformation identification

The images monitored are examined using a manual fact-checking agency and an automated methodology to identify the content misinformation present in the images found in our monitored WhatsApp groups.

5.1.1 Labelling with fact-checking agencies

A list of most shared images was created for the election event and shared with fact-checking agency Lupa [1]in Brazil. We can see from the analysis that, there was a huge number of images containing misinformation during the 2018 Brazilian elections. 36.2% of the images with factual information were identified containing misinformation, whereas 53.2% of them included misleading and inconclusive content, and only 10.6% were verified as true information.

5.1.2 Automatic methodology

The authors have implemented an automated approach to identify misinformation in images. Here they search each image shared on the monitored WhatsApp group on the Web using Google Image search. If in the search result, the returned pages belong to any fact-checking domains, then the fact-checking pages are parsed and the annotation used is automatically extracted to label the images as fake or true. This method was applied over all the images shared in our monitored period and we see that only 2 misinformation images were found for the truck drivers strike dataset and 70 images contained misinformation for election dataset.

[1]https://piaui.folha.uol.com.br/lupa

5.2 Propagation Dynamics

Here the same metric as that used in section 3.2 i.e. lifetimes and burst time is used now to understand the propagation dynamics for images containing misinformation for the election campaign period. From this analysis, we see that around 70% of the images remain in the system for up to 100 hours. Whereas, in 60% of the cases, an image with misinformation has burst time within 100 minutes. Burst times tend to be shorter for misinformation contained images suggesting faster propagation for this content. Also after analyzing for the first appearance of these images, it was found that 20% of the images containing misinformation were shared on web and WhatsApp at the same day, whereas, 35% of them were shared on WhatsApp first, suggesting that WhatsApp is also the main source of misinformation found in the web. We also see that it took less than 6 days for the image with misinformation first shown in WhatsApp to reach the outside web.

6 Conclusion

In this paper, the authors have implemented a detailed analysis of the content shared on publicly available WhatsApp groups for two major events in Brazil. We see that images were the most popular and the ones having most misinformation content. The detailed analysis showcased the types of content, its dissemination and the amount of misinformation found in these contents shared over monitored WhatsApp groups. It was found that images with misinformation are much more often shared first on WhatsApp and then on the Web suggesting that WhatsApp is one of the relevant sources for images with misinformation.

References

Constine 2018
Constine, Josh: *WhatsApp hits 1.5 billion monthly users. $19B? Not so bad.* `https://techcrunch.com/2018/01/31/whatsapp-hits-1-5-billion-monthly-users-19b-not-so-bad/`. Version: 2018. – [Online; posted on 31-Jan-2018]

Garimella & Tyson 2018
Garimella, Kiran; Tyson, Gareth: WhatsApp, Doc? A First Look at WhatsApp Public Group Data. (2018), 04

Monga & Evans 2006
Monga, Vishal; Evans, Brian L.: Perceptual image hashing via feature points: Performance evaluation and tradeoffs. In: *IEEE Transactions on Image Processing* 15 (2006), November, Nr. 11