Image based Misinformation on WhatsApp

Kiran Garimella
Institute for Data, Systems and Society, MIT
garimell@mit.edu

Dean Eckles
Sloan School of Management, MIT
eckles@mit.edu

ABSTRACT

WhatsApp is the most widely used messaging app in the world, with over 1.5 billion active users. Political parties have widely started using WhatsApp for political campaigning with recent media reports claiming that election campaigns by certain political parties have created tens of thousands of WhatsApp groups to reach citizens. Such use by political parties comes with numerous challenges, including the danger of the platform being "weaponized" and used for spreading misinformation and polarizing content.

In this paper, we propose to analyze and characterize image based misinformation being shared on political WhatsApp groups. We identify misinformation by cross referencing images shared on WhatsApp with a large database of fact checked images. Using these images, we try to create a typology of various types of visual misinformation, and understand how visual information is used as a tool of deceit. We also study the spread of such misinformation across other social media platforms and understand the role that WhatsApp plays in such spread. We finally detail the challenges in solving the problem of image based misinformation and the tools and advances required to address the issue.

1 INTRODUCTION

The ever increasing availability of cheap smartphones has spawned a massive surge in new users going online for the first time. This trend is especially stark in developing countries, where hundreds of millions of people have started using the Internet for the first time, primarily through mobile devices. It is estimated that in India alone, at least 400 million users will enter the World Wide Web ecosystem in the next five years.[1] India is adding around 80 million new WhatsApp connections a year, many of them first-time internet users. Hence, and in particular given its penetration on multiple mobile devices and in low connectivity areas, WhatsApp has the potential to empower the next billion users entering the Web.

Political actors have taken up on this wave of new users and started to engage and reach a new potential base through WhatsApp. For example, in a 2017 survey, around one-sixth of WhatsApp users in India said they were members of a group started by a political leader or party.[2] WhatsApp is also being aggressively used for election campaigning, as evidenced in the recent Indian elections, where hundreds of thousands of WhatsApp groups were created to reach the electorate.[3] Political parties use the platform to rally supporters, and dedicated "IT cells" are being setup to use the platform for spreading information and party material, which is then forwarded by their supporters to millions of others, from person to person.[4]

Unfortunately, although conceived as a politically neutral platform, WhatsApp is not immune to the recent spike in fake news, misinformation, and politically polarizing content. This trend is particularly worrying in developing countries like India, where access to third-party independent sources for fact checking is limited, education is a challenge, and cyber literacy is not keeping up with the pace of digitization. Adding to that, for many users, WhatsApp is their first media experience, and they may thus be far more susceptible to manipulation than more seasoned Internet users. We've already observed the grave consequences of such online misinformation leading to real life threats like lynching and social disorder.[5]

Through its design, WhatsApp enables information propagation through multi-media like images, video and audio. It has been empirically observed that image based misinformation is often more viral than text.[6] However, most existing research studying misinformation has been using text, primarily depending on news domains posting fake information. Little is known about how images spread, who spreads them and how we can identify misinformation in non-textual modalities.

Visual misinformation differs from text-based information in important ways, since processing images involves different cognitive processes in our brains. Research on the cognitive process behind processing doctored images [3] shows that humans are bad at identifying doctored images. One recently studied form of image-based misinformation in the US is memes [8, 9]. It has been shown that fringe web platforms like 4chan and Discord are used to create and spread toxic memes that later enter the mainstream media.

In this extended abstract, we present our initial analysis understanding the role of image based misinformation on semi-closed, yet massively popular platforms like WhatsApp. We use a large set of fact checked images from two countries (India and Indonesia) to identify misinformation on WhatsApp. Using this small set of misinformation images, we create a characterization of the types of misinformation being shared on WhatsApp and use that to enable human coding of other images. We also collect data from other mainstream social networks like Twitter, and Facebook and study how misinformation spreads through WhatsApp into these platforms.

Although past studies have investigated WhatsApp usage via methodologies such as interviews [1], we believe it is important that we collect data from public resources at a scale larger than what is possible manually. Hence, we plan to build upon our previous expertise on monitoring public WhatsApp conversations [2], including news sharing, so that we can ground our understanding in hard data. Based on the tools and algorithms, our goal in this project is to get a comprehensive, sociologically rooted understanding of the way image based misinformation spreads on WhatsApp.

[1] http://mashable.com/2016/08/19/india-internet-users-post-pc-era
[2] https://www.livemint.com/Technology/O6DLmIibCCV5luEG9XuJWL/How-widespread-is-WhatsApps-usage-in-India.html
[3] https://www.nytimes.com/2018/05/14/technology/whatsapp-india-elections.html
[4] https://www.newslaundry.com/2017/03/17/how-bjps-it-cell-waged-war-and-won-in-up
[5] https://www.bbc.com/news/world-asia-india-44897714
[6] https://www.poynter.org/fact-checking/2019/these-misleading-images-got-more-engagement-on-faceb

This is not a problem just specific to India or for WhatsApp. Similar characteristics exist in most developing countries where WhatsApp is a dominant player (e.g., Indonesia, Mexico, Brazil). But there is hardly any understanding of the role of WhatsApp in any of these countries. We also expect our results to help understand other closed chat platforms like Discord, Telegram or Signal.

2 DATA

In this section, we detail the various datasets we use in our research. Our analysis relies on large datasets obtained from public political groups on WhatsApp[7] (Section 2.1). To understand the spread of misinformation across platforms, we also collect data from other relevant social media platforms (Section **??**). Finally, to automatically identify images containing misinformation, we also obtain a large number of images from fact checking websites (Section 2.2).

2.1 WhatsApp data

In this research, we use data collected from public WhatsApp groups discussing politics. Public groups on WhatsApp are groups which are open for anyone to join. We collected a large list of public WhatsApp groups via lists publicized on well known websites[8] or social media. Figure 1 shows two public groups advertised by political parties on Twitter.

We joined and obtained data from over 10,000 political groups from India and over 250 groups from Indonesia. From these groups, we obtained all the text messages, images, video and audio shared in the groups. For each group, we also know the members of the group, and the admin(s) of the group. Across the two countries, we collected over 2 million images shared on these political groups.

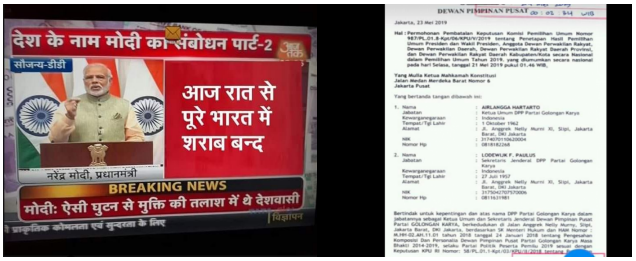
(a) (b)

Figure 1: Examples of official political party accounts sharing their WhatsApp groups on Twitter (Left: BJP, Right: Congress).

2.2 Fact checked images

Finally, in order to create a baseline set of images containing misinformation, we also collected all images which were fact checked on popular fact checking websites from India and Indonesia. This gave us close to 30,000 images. Examples of two fact checked images from India and Indonesia are shown in Figure 2.

Note that not all images which are fact checked contain misinformation. We only used these images as a noisy source of ground truth, i.e., if an image present on a fact checking website is also present in our database, we flag it as a *potential* source of misinformation. These potential sources were then be manually verified to check if they are indeed misinformation or not. This gave us a set of 450 images which were shown to contain misinformation.



(a) (b)

Figure 2: Examples of fact checked misinformation images. (a) A screen grab of a TV program from India stating (falsely) that alcohol is banned, and (b) A picture of a false petition claiming that the opposition wants to cancel elections in Indonesia.

3 CHARACTERIZING IMAGE BASED MISINFORMATION

Based on these rich datasets, we pose the following research questions:

By matching the images from fact checked websites to the millions of images we obtained from WhatsApp using image hashing techniques [5], we can find examples of image based misinformation at scale. By analyzing these examples, we aim to create a typology of the various types of image based misinformation. This will help us understand the different types of deception used. We also plan to use such characterization to help guide manual coding of our images.

Based on manual coding, we were able to identify 4 main types of images based misinformation – (i) old images that are taken out of context and reshared, (ii) memes - funny, yet misleading images, usually containing incorrect stats or quotes, (iii) images which are manipulated/edited, and, (iv) other types of fake images, consisting of health scares, fake alerts, etc. An important aspect to note here is that roughly 15% of the images in our dataset were false health information. Given that the images are from political groups, this is a surprisingly high number. Also, we should note that this is not an exhaustive list of categories and could be specific to our dataset.

Examples of each category of such images is shown in Figure 3.

We find that old images that are taken out of context, also termed cheap or shallow fakes [4] make up roughly **30%** of our misinformation image dataset. The next popular category is the simple doctored/photoshopped images, which make up roughly 20% of the images, followed by fake quotes/stats which make up 10% of the images. Note that just the top three categories of misinformation constitute 60% of fake images.

The characterization of the images presented above helps us in two ways: (1) Can we develop automated techniques to identify misinformation in images?, (2) Can we understand the motives behind posting these misinformation?

[7] Any group on WhatsApp which can be joined using a publicly available link is considered a public group.

[8] For example, https://joinWhatsAppgroup.com/

(a) (b) (c) (d)

Figure 3: Examples of various types of misinformation images. (a) A false quote praising the Indian prime minister, (b) A false health scare, (c) A photoshopped false image, and (d) An image of a plane crash taken out of context.

Tackling these research questions would require advances in both technical as well as labelling techniques. Firstly, manually processing millions of images is not possible. Identifying misinformation automatically using image processing tools, even with the new breed of deep learning based image processing techniques [6], are not up to the task at hand. New computer vision techniques are required for processing and pruning out images to make it tractable for humans to manually annotate images. For instance, we could use automated techniques to match visually similar images using hashing techniques. To this end, we used the state of the art image matching techniques developed and being used at Facebook[9]. The hashing algorithm can detect near similar images, even if cropped differently or have small amounts of text overlaid on them.

Secondly, labelling misinformation on chat platforms is a hard task, due to the importance of context. An image might be considered misinformation depending on the context in which it was shared. An example of the importance of context is shown in Figure 4. As we see, the image can be misinformation or not depending on what was said along with that image. However, in a chat interface, the messages are often not contextualised or verifiable via surrounding links or news stories. Specialized interfaces need to be built to enable labelling of such images, which reconstruct the chat timeline, and at the same time, taking into account the privacy of the users involved in the discussion.

4 AUTOMATED FACT CHECKING

As we saw in Section 3, just three categories of images make up for almost 60% of the misinformation images. This leads us to question if we can develop automated techniques that can help identify at least some of these types of misinformation, to address the problem. Using state of the art hashing techniques, we can easily recover images which are near duplicates. This way, we can identify if an image has been shared in the past and flag images for potential misinformation if shared out of context. However, in practice, these tools are fragile and do not work as expected. Images that are visually similar for a human eye and hard to detect automatically using such matching techniques. Figure 5 shows an example of visually similar images that hashing based techniques can not identify.

The next most popular type of misinformation in our data are photoshopped/manipulated images. Image processing techniques exist to detect manipulated JPEG images [7]. We tried using these

[9]https://github.com/facebook/ThreatExchange/blob/master/hashing/hashing.pdf

(a) Proof that Obama is a muslim. (b) Obama at his half brother's wedding.

Figure 4: Examples of the same image being misinformation and not. The image on the left, along with the caption make it false.

(a) (b)

Figure 5: Examples of similar images that can not be detected using automated techniques. The image on the right is a cropped version of the image on the left, but an automated technique can not detect that they are similar.

techniques to identify if an image has been digitally manipulated. Figure 6 shows two examples. Figure 6a shows image manipulation which is deliberate and is of interest in our case, where as Figure 6b shows a benign meme which is composed of individually photoshopped images. Though this algorithm works as expected to detect image manipulation, most political memes shared in our data are also 'manipulated' to a certain extent.

To conclude, in this section, we tried to automate the fact checking of certain types of image misinformation, but we can see that the state of the art techniques still do not perform up to the mark to do it automatically.

5 WHO IS SHARING THESE IMAGES?

Another important aspect of trying to characterize the misinformation in the form of images is that we can understand who is sharing these images. From our dataset, we observe that over 3,000 users shared at least one of such misinformation images. Over 200 users shared more than 5 misinformation images. We also found that 10%

(a) (b)

Figure 6: Detecting image manipulation automatically. a. An image of the Narendra Modi and the automatic detection of the manipulated parts. b. A benign meme which is also tagged as manipulated by our tool.

of the users who shared these images are admins of their respective groups, indicating potential malicious intent.

It is even more important to study *why* people share such misinformation. From looking at the different types of misinformation images, we can identify the following potential reasons: (i) Low cost of forwarding, (ii) Lack of awareness, (iii) Caring about others[10], and (iv) Malicious intent.

The exact intention is hard to identify from observational data like ours, but we plan to conduct a survey of users in our data using Facebook ads, which might help answer some of these questions [10].

6 CONCLUSIONS

In this work, we analyzed a large scale dataset collected from public political groups on WhatsApp. By matching images shared in our data with publicly available fact checked images, we created a dataset of 450 images which are labelled by fact checkers as misinformation. We manually coded these images and found that just three categories of images – (i) old images taken out of context and posted again, (ii) photoshopped images, and (iii) false quotes and statistics make up roughly 60% of the misinformation in our dataset. We used automated techniques to identify misinformation at scale, and showed that it is not trivial and that the state of the art needs to be improved to be able to be deployed.

Our work provides a first look at misinformation on WhatsApp, through images. Given that a large majority of the messages shared on WhatsApp consitute images, understanding the prevalence of misinformation on such modalities is an important task. Our data only constitutes publicly shared WhatsApp messages, and hence is subject to various biases. However, the characterization of the types of images helps us understand the underlying motivations and potential methods to prevent the spread of such misinformation. Ultimately, since WhatsApp is a closed platform, any technique that is developed to detect and debunk misinformation can only be implemented by WhatsApp/Facebook, which should build tools necessary to help users identify misinformation effectively.

[10]This is especially true in case of health related misinformation or child kidnapping rumors where users just forward content even if they are not sure if it is true, just in case it helps others.

7 AUTHOR BIOGRAPHY

Kiran Garimella is the Michael Hammer postdoctoral fellow at the Institute for Data, Society and Systems (IDSS) at MIT. Before joining MIT, he was a postdoc at EPFL, Switzerland. His research focuses on using digital data for social good, including areas like polarization, misinformation and human migration. His work on studying and mitigating polarization on social media won the best student paper awards at WSDM 2017 and WebScience 2017. Kiran received his PhD at Aalto University, Finland, and Masters & Bachelors from IIIT Hyderabad, India. Prior to his PhD, he worked as a Research Engineer at Yahoo Research, Barcelona, and QCRI, Doha.

Dean Eckles is a social scientist and statistician. Dean is the KDD Career Development Professor in Communications and Technology at Massachusetts Institute of Technology (MIT), an assistant professor in the MIT Sloan School of Management, and affiliated faculty at the MIT Institute for Data, Society & Systems. He was previously a member of the Core Data Science team at Facebook. Much of his research examines how interactive technologies affect human behavior by mediating, amplifying, and directing social influence – and statistical methods to study these processes. Dean's empirical work uses large field experiments and observational studies. His research appears in the Proceedings of the National Academy of Sciences and other peer-reviewed journals and proceedings in statistics, computer science, and marketing. Dean holds degrees from Stanford University in philosophy (BA), symbolic systems (BS, MS), statistics (MS), and communication (PhD).

REFERENCES

[1] Karen Church and Rodrigo de Oliveira. 2013. What's up with whatsapp?: comparing mobile instant messaging behaviors with traditional SMS. In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services*. ACM, 352–361.

[2] Kiran Garimella and Gareth Tyson. 2018. WhatsApp, Doc? A First Look at WhatsApp Public Group Data. In *International AAAI Conference on Web and Social Media*. 511–518.

[3] Sophie J. Nightingale, Kimberley A. Wade, and Derrick G. Watson. 2017. Can people identify original and manipulated photos of real-world scenes? *Cognitive Research: Principles and Implications* 2, 1 (18 Jul 2017), 30. DOI: https://doi.org/10.1186/s41235-017-0067-2

[4] Britt Paris and Joan Donovan. 2019. Deepfakes and cheap fakes: the manipulation of audio and visual evidence. *Data and Society* (2019).

[5] Ramarathnam Venkatesan, S-M Koon, Mariusz H Jakubowski, and Pierre Moulin. 2000. Robust image hashing. In *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, Vol. 3. IEEE, 664–666.

[6] Ren Wu, Shengen Yan, Yi Shan, Qingqing Dang, and Gang Sun. 2015. Deep image: Scaling up image recognition. *arXiv preprint arXiv:1501.02876* (2015).

[7] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. 2019. ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries With Anomalous Features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9543–9552.

[8] Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. 2019. Characterizing the Use of Images by State-Sponsored Troll Accounts on Twitter. *arXiv preprint arXiv:1901.05997* (2019).

[9] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*. ACM, 188–202.

[10] Baobao Zhang, Matto Mildenberger, Peter D Howe, Jennifer Marlon, Seth A Rosenthal, and Anthony Leiserowitz. 2018. Quota sampling using Facebook advertisements. *Political Science Research and Methods* (2018), 1–7.