# Considerations in Archiving Misinformation from Encrypted Messaging Apps

Authors: Tarunima Prabhakar, Denny George
Affiliation: Tattle

Date: 1st Oct 2019

## Introduction

A year after Brexit and US election brought the spotlight on social media platforms for their role in amplifying misinformation, WhatsApp came into prominence as a vector for rumours in the global South. While misinformation had long been a feature in groups conversations on the platform, the spate of lynchings in India triggered by WhatsApp rumours[i] brought the platform to global attention.

The surge of messaging apps such as TikTok and ShareChat; and Facebook's announcement of plans to merge WhatsApp, Facebook and Instagram chat[ii] has come with a recognition that misinformation on encrypted messaging platforms could soon be a global phenomenon.[iii]

Tattle is a civic tech project emanating from India that aims to make verified information more easily accessible to mobile first users, in languages that they are comfortable with. Started with the intent of addressing misinformation on WhatsApp, the project has expanded in scope to address misinformation on chat apps and encrypted networks in general.

In this presentation I will present some of the challenges and associated design decisions in implementing one of Tattle's goals[1]- the creation of a globally accessible archive of multi-media messages circulated on chat apps.

## The Curious Case of WhatsApp

[1] Tattle is also building open source tools that let users verify the content they receive via these archives: https://gettattle.app/

In India, WhatsApp's rising popularity has mirrored a rapid increase in mobile phone and Internet penetration[iv]. The ease of creating and sharing audio-visual content has made the platform accessible even to users with minimal traditional and digital literacy.

As an encrypted platform, WhatsApp cannot access content circulating on it. Consequently, vis-à-vis other social media networks, there is lower information asymmetry between the company and the public. Encryption however, also reduces the agency of the company in removing problematic content- WhatsApp cannot automate detection and content removal, as its parent company Facebook does[v], without weakening encryption standards.

There are other aspects that make misinformation on WhatsApp distinct from other platforms. First, misinformation is often circulated on closed groups where membership is contingent on invitation[vi]. Secondly, WhatsApp does not allow discovery of new connections via the app itself[2]. Finally, unlike Facebook or YouTube, there is no algorithmic curation of content on the platform. Sharing by individuals is the *only* contributor to virality of content on the platform.

These aspects make content discovery and tracing source of origin of a particular message challenging. This not only raises difficulties for fact checkers responding to online misinformation, but also for those trying to understand information networks on the platform. Several questions such as 'how quickly and how widely does content travel on the platform? Do WhatsApp videos have a shelf life or are they recycled in newer forms?' remain unanswered.

An archive of content circulated on WhatsApp, and associated metadata such as time of receiving the message, addresses some of these issues. For fact checkers, it provides a database against which they can compare content received for verification. For researchers, it is a source of content and temporal analysis. But an archive is not neutral. It exerts power in what it chooses to include and exclude.[vii] It also exerts power through its mode of management and rules of access.

## Data Acquisition: The Greedy Approach

Over the last two years, several researchers and fact checking groups have collected content circulated on WhatsApp for different purposes. While most researchers have focused on scraping public WhatsApp groups[viii,ix] fact checking groups source content from a number of channels, including emails and fact checking helplines run over WhatsApp.[x]

The lynchings in India highlight the urgency of timely detection and response. With the intent of surfacing viral content as quickly as possible, Tattle has adopted a greedy approach to data collection. Besides crowdsourcing data via an android app, Tattle is working on automating group discovery and content extraction from WhatsApp to the extent possible.

## Open Challenges

[2] Users must have access to another user's phone number to find them on WhatsApp.

## Ethics of Data Collection from Public WhatsApp Groups

At present there are no guidelines for ethical data collection on WhatsApp. Every research group has self-defined a configuration of consent framework and data sharing practices. For example, Narayanan et al. [ix] chose to declare their presence and intent in every public group they joined and as a result were removed from some of them. Garimella et al.[viii] on the other hand did not declare their presence and intent and have released a limited dataset.

At Tattle, we do not declare our intent when we join public WhatsApp groups. Our working assumption is that public WhatsApp groups are an exception to otherwise private communication on the platform and are made public with the explicit intent of being discoverable. This is an assumption that we will revisit as the project evolves.

## Flagging Problematic Content

Not all collected data can be opened to the public. A casual google search for public WhatsApp groups results in lists of 'Adult' WhatsApp groups[xi]. In greedy scraping of WhatsApp groups, we often encounter pornographic and violent content. For example, groups launched under a popular theme of 'Jobs' are often appropriated for pornographic content. Such content cannot be shared on a public archive and must be flagged soon after collection.

Owing to the volume of data scraped, Tattle will use machine learning based approaches to flag problematic content. As all ML systems, this classification system too will have prediction errors and there is unavoidable subjectivity in how the algorithms are optimized for different error rates. A choice to minimize false negatives (violent content not getting flagged) may come at the cost of higher false positives (non-violent content getting flagged as violent content). Furthermore, there can be disagreements on what expression is problematic and therefore not suitable for a public archive. These norms are likely to be culture and context specific.

How best to incorporate the preferences of multiple stakeholders in what should or should not be opened, is a question that Tattle as any open archive must address.

## Unwanted spotlight on previously obscure content [xii,xiv]

With an online open access archive it is difficult to predict or control the ways in which the data will be used. One concern is that by surfacing content circulating in different geographies, the archive will become a source for ill-intended content creators, confounding the problem the archive is directed to solve.

## Access and Use

Misinformation on WhatsApp is often hyperlocal.[xiii] While local fact checkers and civil society actors are best placed to act locally, they might not have the institutional support required for academic-industry partnerships through which social media data is shared. Even in newer research collaborations, access to data remains privileged[xiv,xv]. An open data archive circumvents gatekeeping costs, minimizes monetary barriers to access and also enables timeliness in access of data.[3] Given that misinformation is often globally sourced but locally contextualized[xvi], a global archive of viral content on social media can help shave off critical time in local fact checking.

We recognize that while not all data collected can be made public, it may still be useful for research. Preserving some data for restricted access has practical applications, though it raises questions around how these rights are managed. Aronson notes, "…the ethics of providing access to archival materials is a 'thorny problem.' It is not a one and done policy decision, instead requires constant deliberation and negotiation."[xvii]

## Precedents for Open Archives

### Una Hakika

The Sentinel Project launched Una Hakika in 2014 in the Tana Delta region in Kenya. The region is especially susceptible to ethnic conflict due to misinformation. "…the majority of Tana Delta residents consider themselves to be well-informed about their own villages and national events in Kenya as a whole, but not about events in neighbouring villages (which may be only a few hundred meters away or on the other side of a river) or their county. This gap is particularly dangerous when the residents of those neighbouring villages are members of another ethnic group that is seen as hostile." [xviii] The region is one of the least developed in Kenya but has high mobile phone and Internet usage. The project uses SMS[4] to collect rumours circulating in the region and to provide feedback on the rumours. Community leaders play a pivotal role is disseminating verified stories.

The rumours and verification status is shared on a pulbic 'wikirumours' database[5]. The database allows aggregation of multiple sightings of the same rumour which helps in prioritizing content for verification. The project has been replicated in other conflict regions in East and Central Africa.

### Wayback Machine (Archive.org)

The Internet Archive is a US based non-profit with a stated mission to 'provide Universal Access to All Knowledge'[xix]. Amongst other digital artifacts, Archive.org has been

[3] As the SAA Code of Ethics for Archivists notes, "use is the fundamental reason for keeping archives… Archivists actively promote open and equitable access to the records in their care." https://www2.archivists.org/statements/saa-core-values-statement-and-code-of-ethics

[4] Also known as text message.

[5] https://www.unahakika.org/search_results/report%3Dcommon

cataloguing web pages since 1996. The webpages are accessible to the public through the archive's online portal, the Wayback Machine. While most of the webpages are made public, there have been some notable exceptions where archived pages have been deleted[xx]. As with other archival initiatives, the Internet Archive has to contend with subjectivity in editorial choices[xxi]. The archive's users have also highlighted it as an enduring source of copyrighted or extremist content that might have been removed from the original website[xxii]. The archived webpages have been used in journalistic investigations[xxiii] and litigations[xxiv] involving contestations in current and historical online activity by an individual or organization[xxv].

# Discussion

Archival initiatives such as The Wayback Machine and Una Hakika exemplify the utility as well as challenges associated with archiving content in the digital age. The medium of communication is a critical consideration in archival process. It influences not only what is archived, but how the archived data is opened and shared with public.

Encrypted chat apps such as WhatsApp present novel challenges for archiving. The lack of attributability, communication of related content in separate messages in a thread and multi-media messages contribute to the complexity of Tattle's archival process. In the Indian context, multi-lingual communication adds to the technical complexity of creating a searchable archive.

The ethical challenges raised here are not easily resolved. Indeed, they have much in common with the tensions, increasingly highlighted, in centralized social media platforms such as that between individual consent and transparency; access to information and individual safety; centralization and access. Archiving content on chat apps, however, presents an opportunity for alternate balancing of such tensions. Such an archive might also emerge as a practical aid in understanding and mitigating misinformation on such platforms, without circumventing encryption.

# References

[i] Lynchings Due To Whatsapp Rumours Claim 19 Lives in Two Months. The Quint. 17.07.18. Accessed May 31, 2019. https://www.thequint.com/news/india/lynchings-due-to-rumours-spreading-on-whatsapp

[ii] Isaac, Mike. "Zuckerberg Plans to Integrate WhatsApp, Instagram and Facebook Messenger." The New York Times. January 25, 2019. Accessed June 01, 2019. https://www.nytimes.com/2019/01/25/technology/facebook-instagram-whatsapp-messenger.html.

[iii] De, Anamitra. "Can We Trust Facebook to Keep Our "Digital Living Rooms" Safe From Liars, Racists, and Haters? | Omidyar Network." Omidyar Network Blog. March 11, 2019. Accessed June 01, 2019. https://www.omidyar.com/blog/can-we-trust-facebook-keep-our-digital-living-rooms-safe-liars-racists-and-haters.

[iv] Press Trust of India. "Internet Users in India To Reach 627 Million in 2019". Press Trust of India. Mar 06, 2019. Accessed May 31, 2019. https://telecom.economictimes.indiatimes.com/news/internet-users-in-india-to-reach-627-million-in-2019-report/68297051.

[v]Vincent, James. "Facebook Is Using Machine Learning to Spot Hoax Articles Shared by Spammers." The Verge. June 21, 2018. Accessed May 31, 2019. https://www.theverge.com/2018/6/21/17488040/facebook-machine-learning-spot-hoax-articles-spammers.

[vi] Choi, Boreum, and Inseong Lee. "Trust in Open versus Closed Social Media: The Relative Influence of User- and Marketer-Generated Content in Social Network Services on Customer Trust." *Telematics and Informatics* 34, no. 5 (August 1, 2017): 550–59. https://doi.org/10.1016/j.tele.2016.11.005.

[vii] Carter, Rodney G. S. "Of Things Said and Unsaid: Power, Archival Silences, and Power in Silence." *Archivaria* 61, no. 0 (September 25, 2006): 215–33.

[viii] Garimella, Kiran, and Gareth Tyson. 2018. "WhatsApp, Doc? A First Look at WhatsApp Public Group Data." *ArXiv:1804.01473 [Cs]*, April. http://arxiv.org/abs/1804.01473.

[ix]Narayanan, Vidya. Kollanyi, Bence. Hajela, Ruchi. Barthwal, Ankita. Marchal, Nahema. Howard, Philip. "News and Information over Facebook and WhatsApp during the Indian Election Campaign." Data Memo 2019.2. Oxford, UK: Project on Computational Propaganda. comprop.oii.ox.ac.uk

[x] Lomas, Natasha. "WhatsApp Adds a Tip-line for Gathering Fakes Ahead of India's Elections – TechCrunch." TechCrunch. April 02, 2019. Accessed May 31, 2019. https://techcrunch.com/2019/04/02/whatsapp-adds-a-tip-line-for-checking-fakes-in-india-ahead-of-elections/.

[xi] https://www.whatsappgrouplinks.org

[xii] Freelon, Deen et al., Beyond the Hashtags: #Ferguson, #Blacklivesmatter, and the Online Struggle for Offline Justice
(Washington, DC: Center for Media and Social Impact, 2016), 86.

[xiii] "BJP's Social Media 'Yodha' from Cooch Behar Is an Admin of over 1,000 WhatsApp Groups." Moneycontrol. Apr12,2019. Accessed May 31, 2019. https://www.moneycontrol.com/news/trends/bjps-social-media-yodha-from-cooch-behar-is-an-admin-of-over-1000-whatsapp-groups-3814131.html. Accessed May 26, 2019.

[xiv] https://www.whatsapp.com/research/awards/

[xv] https://socialscience.one/

[xvi] Kumar, Aishwarya. "Death By WhatsApp: How a Video Shot in Pakistan Led to Death of 30 People in India." News18. July 09, 2018. Accessed May 31, 2019. https://www.news18.com/news/india/death-by-whatsapp-how-a-video-shot-in-pakistan-led-to-death-of-30-people-in-india-1805733.html.

[xvii] Aronson, Jay D. (2017) "Preserving Human Rights Media" Genocide Studies and Prevention: An International Journal: Vol. 11: Iss. 1: 82-99. Accessed May 31, 2019. https://scholarcommons.usf.edu/gsp/vol11/iss1/9/.

[xviii] https://thesentinelproject.org/2014/02/17/how-it-works-una-hakika/

[xix] https://archive.org/about/

[xx] Nelson, Steven. "Wayback Machine Won't Censor Archive for Taste, Director Says After Olympics Article Scrubbed". US News. Aug 17, 2016. Accessed September 15, 2019. https://www.usnews.com/news/articles/2016-08-17/wayback-machine-wont-censor-archive-for-taste-director-says-after-olympics-article-scrubbed

[xxi] Lafrance, Adrienne. "Raider of the Lost Web". The Atlantic. Oct 14 2015. https://www.theatlantic.com/technology/archive/2015/10/raiders-of-the-lost-web/409210/

[xxii] https://archive.org/post/931527/atchiveorg-is-widely-used-by-terrorists-do-you-know-that

[xxiii]Lepore, Jill. "The Cobweb". Annals of Technology, The New Yorker. Jan 19 2015. Accessed September 15, 2019. https://www.newyorker.com/magazine/2015/01/26/cobweb

[xxiv] https://casetext.com/case/telewizja-polska-usa-4

[xxv]Hodgson, Camilla. "How the Internet Archive is waging war on misinformation". Financial Times. Sep 16 2019. Accessed Sep 20 2019. https://www.ft.com/content/5be1f2ee-d60b-11e9-a0bd-ab8ec6435630