Tools for Countering Misinformation on Encrypted Chat Apps

Data Collection and Management at Tattle

Swair Shah
Data Scientist
Tattle
swairshah@gmail.com

Denny George
Design Technologist
Tattle
denny@nyu.edu

Tarunima Prabhakar
Research Associate
CLTC, UC, Berkeley
tarunima@berkeley.edu

1 Introduction

Messaging apps like WhatsApp are a prominent platform for misinformation in the Global South. Facebook's announcement of plans to integrate its distinct messaging services (Instagram, Messenger, WhatsApp) has come with a recognition that misinformation on chat apps could become a global phenomenon. Tattle[1] is a civic tech project working to create a globally accessible archive of multimedia messages circulated on chat apps. Such an archive is valuable for fact checking groups, as well as for researchers trying to understand information networks on these platforms. In this talk we will share Tattle's approach to the challenge of data collection from WhatsApp; and our approach to indexing and querying multilingual and multimedia content.

2 Misinformation- The WhatsApp Challenge

WhatsApp is distinct from social networks in several critical ways. First, on WhatsApp, content is often circulated on closed groups whose membership is contingent on invitation. Second, WhatsApp does not allow discovery of new connections via the app itself. Third, private sharing leads to isolated, hyper-local networks. These factors make it harder to understand information networks on the platform or identify viral content.

3 Existing Approaches

The need for extracting content from WhatsApp is felt by fact checkers and researchers alike. Currently fact checking groups rely on crowdsourcing from concerned individuals through WhatsApp 'tiplines', emails and social media handles, often responding to requests for verification of the same content multiple times. After verification, the fact checked content is shared on the fact checking group's website and social media handles.

Researchers aiming to understand information networks on WhatsApp have typically used (Garimella and Tyson, 2018) approach of joining multiple public WhatsApp groups and decrypting the database to extract content shared on these groups. Several factors such as high volume of problematic content and uncertainty around use of such content have led researchers to restrict access to data collected (Prabhakar, 2019).

4 Tattle's Approach

4.1 Data Collection

In addition to collecting data by a crowdsourcing app and decrypting local WhatsApp database (Garimella and Tyson, 2018), we are also using Android's Accessibility Service API to scrape public whatsApp groups.

4.1.1 Android Accessibility Service Approach

The service provides APIs (Google, 2000) to build apps that assist users with disabilities. Android exposes a tree of AccessibilityNodeInfo objects, a class that represents a user interface node as well as actions that can be requested from it. In the case of WhatsApp this includes accessing individual messages and the few buttons in the chatroom interface. The service also provides functionality to mimic user actions such as clicking and swiping. A combination of these two methods makes it possible to automate the process of forwarding incoming WhatsApp messages to a Tattle server.

This method is versatile and can be extended to other closed chat apps such as Viber, Telegram. This approach has high potential for misuse. Thus we have not opened this service yet, and are delib-

[1] More information about Tattle can be found at https://gettattle.app/

erating about safeguards that are necessary when sharing or advocating this method for data collection.

4.2 Content Search

All data collected through the aforementioned approaches is indexed and added to a database. The database can be queried for exact and approximate matches of a specific multimedia message. Tattle's search service handles three kinds of content searches:

- **Exact document search** or duplicate detection is used to find the previous occurrences of the exact same document.
- **Approximate search** deals with cases where a document is similar but not a duplicate of another document. For example it may contain extra information or omit some details. This is helpful in tracing content that is a cropped from another image, has additional filters for effect or intentional distortion or images where a text snippet has been added to create context.
- **Related content search** is useful to find content of a particular topic. Such search would be helpful, for example, in retrieving different content linked to a similar event.

The search service computes fingerprint for the query and computes similarity using euclidean norm. In order to speed up the operations, we rely on approximate nearest neighbor methods (Indyk and Motwani, 1998; Bernhardsson, 2018). We now describe our approaches for computing fingerprints for different types of documents.

4.2.1 Images

The image fingerprint method adopted uses a ResNet (He et al., 2016) Convolutional Neural Network with 18 layers trained in the ImageNet dataset (Deng et al., 2009). The image is passed through ResNet-18. Subsequently, the feature vector from the last fully connected layer is used as an image fingerprint. This approach works well for approximate search and can also be applied to exact search.

4.2.2 Text Documents

For each text document we compute document fingerprints. We first compute word vectors for each word using methods described in (Mikolov et al., 2013). A document fingerprint is calculated as the average of generated word vectors. Besides English, we support embeddings for multiple Indian languages using embeddings provided in (Grave et al., 2018). Like image fingerprints, document fingerprints facilitate exact and approximate search.

4.2.3 Images with Text

In many cases the images indexed by Tattle have large portions of text in them. For such documents, we extract the text from the images using Google Vision API[2] which supports multiple Indian Languages. This results in a hybrid document with two fingerprints corresponding to the image and extracted text. The overall document fingerprint in this case is a concatenation of the image and text fingerprints.

4.2.4 Planned Approaches

We have been experimenting with developing bespoke methods to fingerprint and search for the type of videos that are found on WhatsApp. We are also experimenting with using (Le and Mikolov, 2014) to generate document embeddings instead of taking an average of the word embeddings in the document.

4.3 Access and Use

Misinformation on WhatsApp is often hyperlocal (Bose, 2019). While local fact checkers and civil society actors are best placed to act locally, they might not have the institutional support required for academic or industry partnerships through which data is often made available. An open data archive circumvents gate-keeping costs, minimizes monetary barriers to access and enables timeliness in access of data. We recognize that not all data collected can be made public and are exploring strategies for flagging problematic content that should not be made open.

We are elaborating on the ethical considerations in data collection from closed networks in a separate research project. (Prabhakar, 2019)

5 Discussion

In our talk we will elaborate on Tattle's approaches in context of peculiarities of information networks on WhatsApp in India; and how these approaches can be extended to other chat platforms.

[2] https://cloud.google.com/vision/docs/ocr

References

Erik Bernhardsson. 2018. *Annoy: Approximate Nearest Neighbors in C++/Python*. Python package version 1.13.0.

Ankita Bose. 2019. Bjps social media yodha from cooch behar is an admin of over 1,000 whatsapp groups. https://bit.ly/2Kku9db.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Kiran Garimella and Gareth Tyson. 2018. Whatapp doc? a first look at whatsapp public group data. In *Twelfth International AAAI Conference on Web and Social Media*.

Google. 2000. Accessibility service api documentation. https://developer.android.com/reference/android/accessibilityservice/AccessibilityService.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Piotr Indyk and Rajeev Motwani. 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613. ACM.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tarunima Prabhakar. 2019. Considerations in archiving content from encrypted messaging apps. http://blog.tattle.co.in/considerations-in-archiving\-misinformation-fromencrypted-\messaging-apps.