Harvard Kennedy School Misinformation Review[1]
January 2022, Volume 3, Issue 1

DOI: https://doi.org/10.37016/mr-2020-91
Website: misinforeview.hks.harvard.edu

Research Note

Research note: Tiplines to uncover misinformation on encrypted platforms: A case study of the 2019 Indian general election on WhatsApp

There is currently no easy way to discover potentially problematic content on WhatsApp and other end-to-end encrypted platforms at scale. In this paper, we analyze the usefulness of a crowd-sourced tipline through which users can submit content ("tips") that they want fact-checked. We compared the tips sent to a WhatsApp tipline run during the 2019 Indian general election with the messages circulating in large, public groups on WhatsApp and other social media platforms during the same period. We found that tiplines are a very useful lens into WhatsApp conversations: a significant fraction of messages and images sent to the tipline match with the content being shared on public WhatsApp groups and other social media. Our analysis also shows that tiplines cover the most popular content well, and a majority of such content is often shared to the tipline before appearing in large, public WhatsApp groups. Overall, our findings suggest tiplines can be an effective source for discovering potentially misleading content.

Authors: Ashkan Kazemi (1,4), Kiran Garimella (2), Gautam Kishore Shahi (3), Devin Gaffney (4), Scott A. Hale (4,5)
Affiliations: (1) Computer Science & Engineering, University of Michigan, USA, (2) School of Communication and Information, Rutgers University, USA, (3) Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen, Germany, (4) Meedan, USA, (5) Oxford Internet Institute, University of Oxford, UK
How to cite: Kazemi, A., Garimella, K., Shahi, G. K., Gaffney, D., & Hale, S. A. (2022). Research note: Tiplines to uncover misinformation on encrypted platforms: A case study of the 2019 Indian General Election on WhatsApp. *Harvard Kennedy School (HKS) Misinformation Review*, *3*(1).
Received: September 15th, 2021. Accepted: December 17th, 2021. Published: January 31st, 2022.

Research questions

- How effective are tiplines for identifying potentially misleading content on encrypted social media platforms?
- What content is submitted to tiplines for fact-checking?

Essay summary

- A *tipline* is a dedicated service to which messages ("tips") can be submitted by users. On WhatsApp, a tipline would be a phone number to which WhatsApp users can forward information they see in order to have it evaluated by fact checkers.

[1] *A publication of the Shorenstein Center on Media, Politics and Public Policy at Harvard University, John F. Kennedy School of Government.*

- Using state-of-the-art text and image matching techniques, we compared content sent to the tipline to the content collected from a large-scale crawl of public WhatsApp groups (these are WhatsApp groups where the link to join is shared openly), ShareChat (a popular image sharing platform in India similar to Instagram), and fact checks published during the same time in order to understand the overlap between these sources.
- The tipline covers a significant portion of popular content: 67% of images and 23% of text messages shared more than 100 times in public WhatsApp groups appeared on the tipline.
- We found that a majority of the viral content spreading on WhatsApp public groups and on ShareChat was shared on the WhatsApp tipline before appearing in the public groups or on ShareChat.
- Compared to content by popular fact-checking organizations, the messages from tiplines cover a much higher proportion of WhatsApp public group messages. We suspect this is because fact-checking organizations typically fact-check content primarily based on signals from open social media platforms like Facebook and Twitter, whereas the tipline is a crowdsourced collection of content native to WhatsApp.

Implications

Platforms such as WhatsApp that offer end-to-end encrypted messaging face challenges in applying existing content moderation methodologies. End-to-end encryption does not allow the platform owner to view content. Rather, only the sender and recipients have access to the content—unless it is flagged by a receiving user (Elkind et al., 2021). Even though WhatsApp is extremely popular, used by over 2 billion users all over the world, there is currently no large-scale way to understand and debunk misinformation spreading on the platform. Given the real-life consequences of misinformation (Arun, 2019) and the increasing number of end-to-end encrypted platforms, developing tools to understand and uncover misinformation on these platforms is a pressing concern.

One potential solution is to make use of misinformation "tiplines" to identify potentially misleading or otherwise problematic content (Meedan, 2020). A *tipline* is a dedicated service to which "tips" can be submitted by users. On WhatsApp, a tipline would be a phone number to which WhatsApp users can forward potential misinformation they see in order to have it fact-checked. We call the messages sent by users "tips."

While our paper is, to the best of our knowledge, the first peer-reviewed study on WhatsApp tiplines, tiplines are quite common in practice. WhatsApp, for instance, currently lists 54 fact-checking organizations with accounts on its platform.[2] Other efforts include the Comprova project[3] and FactsFirstPH,[4] an initiative of over 100 organizations uniting around the 2022 Philippine presidential election. Tiplines are similar to features on platforms such as Twitter and Facebook that allow users to flag potential misinformation for review, but tiplines are operated by third parties and can provide instantaneous results for already fact-checked claims (Kazemi et al., 2021).

In this study, we used data from a WhatsApp tipline that ran during the 2019 Indian general election as part of the Checkpoint project.[5] Checkpoint was a research project led by PROTO[6] and Pop-Up

[2] https://web.archive.org/web/20211130214745/https://faq.whatsapp.com/general/ifcn-fact-checking-organizations-on-whatsapp/?lang=en
[3] https://firstdraftnews.org/tackling/comprova/
[4] https://factsfirst.ph/
[5] https://www.checkpoint.pro.to/
[6] PROTO is an Indian organization that describes itself as, "a social enterprise that is trying to achieve better outcomes in civic media through collaboration and research" (https://www.pro.to/about/index.html).

Newsroom, technically assisted by WhatsApp.[7] The goal of this project was to study the misinformation phenomenon at scale—natively in WhatsApp—during the Indian general election. The tipline was advertised in the national and international press during the election (e.g., Lomas, 2019). There was an advertising campaign on Facebook, but no specific call to action was present in WhatsApp itself. Table 1 presents some examples of text messages submitted to the tipline. The goal of this article is to understand what content is submitted, analyze how effective tiplines can be for discovering content to fact-check, and shed light on the otherwise black-box nature of content spreading on WhatsApp.

Table 1. *Examples of English text messages forwarded to the WhatsApp tipline to be fact-checked.*

UNESCO Declare India's "Jana Gana Mana" the World's Best National Anthem
When you reach poling booth and find that your name is not in voter list, just show your Aadhar card or voter ID and ask for "challenge vote" under section 49A and cast your vote. If you find that someone has already cast your vote, then ask for "tender vote" and cast your vote. If any polling booth records more than 14% tender votes, repolling will be conducted in such poling booth. Please share this very important message with maximum groups and friends as everyone should aware of their right to vote.
Happened today on 47 street (Diamond Market) New York $100,000 given away in ref to Modi victory .. see how this millionaire Indian is doing ..
Coal India is on the verge of ruin! 85,000 crore loss due to Modi! <url>

Note: Grammar and spelling errors are in the originals. The content we analyzed includes messages in multiple languages and formats (e.g., text, images, and links).

Our results show the effectiveness of tiplines in content discovery for fact-checking on encrypted platforms. We show that:

1. A majority of the viral content spreading on WhatsApp public groups and on ShareChat was shared on the WhatsApp tipline first, which is important as early identification of misinformation is an essential element of an effective fact-checking pipeline given how quickly rumors can spread (Vosoughi et al., 2018).
2. The tipline covers a significant portion of popular content: 67% of images and 23% of text messages shared more than 100 times in public WhatsApp groups appeared on the tipline.
3. Compared to content from popular fact-checking organizations, the messages sent to tiplines cover a much higher proportion of WhatsApp public group messages. While misinformation often flows between platforms (Resende et al., 2019), this suggests that tiplines can capture unique content within WhatsApp that is not surfaced by fact-checking efforts relying on platforms without end-to-end encryption.

These insights demonstrate tiplines can be an effective privacy-preserving, opt-in solution to identify potentially misleading information for fact-checking on WhatsApp and other end-to-end encrypted platforms. At the same time, there is the possibility of malicious uses and attacks on tiplines that may negatively affect fact checkers, share personal information from others, or poison the dataset. As we discuss in the findings, it is necessary to filter spam and other low-quality submissions. We analyzed submissions qualitatively to identify those with a claim that could be fact-checked, but there are several machine-learning approaches in development for this task (e.g., Hassan et al., 2015; Shaar et al., 2021).

[7] Pop-Up Newsroom is a joint project of Meedan and Fathm that designs and leads global election and event monitoring journalism efforts.

Tiplines, like systems for content moderation, must prioritize fact checkers' mental health (Lo, 2020). The Meedan software used in the Checkpoint project, for instance, now uses Google's SafeSearch API to place a content screen over potentially explicit images. Similar systems, however, are needed to protect fact checkers from vicarious trauma as well as personal attacks in audio, video, and text in the myriad languages in which fact checkers operate. We can further reduce harm and malicious activity by designing friction into tiplines such as menu systems and limits on the number of requests per user to prevent denial of service attacks. We are currently investigating the data governance and safeguards needed to share tipline data more widely with academics for research (Meedan, 2021).

In addition to the general public, we see three main stakeholders who could benefit from this research: academics, fact-checking organizations, and social media companies. Researchers or journalists trying to use data from encrypted social media apps like WhatsApp could make use of data from such tiplines to study WhatsApp. The current model for identifying and fact-checking viral content on WhatsApp is to monitor conversations in a convenience sample of public WhatsApp groups (Garimella & Eckles, 2020; Melo et al., 2019). However, this requires technical skill and is resource intensive to manage. To our knowledge, monitoring of public groups has occurred only in academic settings.

Another solution that fact-checking organizations follow is to monitor non-encrypted social media platforms such as Facebook or Twitter and assume that content viral on one of these platforms likely overlaps with viral content on other platforms. Our work shows that there are far more matches between tipline content and public group messages on WhatsApp than between public group messages and either published fact checks or open social media content. This notable difference in the coverage of WhatsApp public groups stresses the opportunity tiplines provide for identifying misinformation on encrypted platforms. Although the volume of messages sent to the tipline is only 10% the volume of messages in the public groups, our analysis shows that tiplines can effectively help discover the most viral content being shared in the public groups. As end-to-end encryption prevents other forms of monitoring, identifying the most popular content on an end-to-end encrypted platform is useful to fact checkers, even if only a subset of that content is actual misinformation. The data we have for analysis does not include the fact-checks for the content submitted to the tipline, but our analysis shows that the majority of content submitted to the tipline contains claims that can be fact-checked.

Further research is needed to determine the best way fact checkers can prioritize content submitted to tiplines, filter spam and low-quality materials, combine signals from other platforms (e.g., from CrowdTangle and/or Twitter), and study the impact of fact-checks distributed via tiplines. Some methods, such as claim extraction (Hassan et al., 2015; Shaar et al., 2021) and claim matching (Kazemi et al., 2021; Shaar et al., 2020), are directly applicable to tiplines, while other aspects require further work. Our analysis shows content is often submitted to tiplines before spreading in larger groups; however, this is only one step of the fact-checking progress. To be effective, we need systems that help fact checkers prioritize content for fact-checking, respond to that content, and disseminate fact-checks before the problematic content spreads widely. Nakov et al. (2021) provide an overview of many ways in which further research and tool development could assist human fact checkers, and nearly all of these are applicable to tiplines as well. Our data predates the introduction of the "frequently-forwarded" flag on WhatsApp, but a report from Spanish fact-checking organization Maldita.es suggests this flag can be very useful for prioritizing content from WhatsApp tiplines (Maldita.es, 2021).

Our analysis also found that most users sending content to the tipline were motivated to have the content they sent fact-checked: users would often follow up on content they submitted if it had not yet been fact-checked. We are unaware of any successful tiplines run solely as research projects, which suggests that fact-checking organizations and academics will need to partner together to scale tiplines and create meaningful tipline experiences for users. This will involve setup costs and take time to foster dedicated contributors who are willing to forward potentially misleading content to a tipline.

It's worth noting that the tipline, public group, and fact-check content we studied were drawn from a specific period of time around a large political event (the 2019 Indian general election). It is unclear how the dynamics would differ for a less eventful time period. Several always-on WhatsApp misinformation tiplines were launched in December 2019, and the number has grown since. We encourage researchers to support civil society organizations running these tiplines, as they represent a valuable way to better understand the dynamics of misinformation on such end-to-end encrypted platforms.

Tiplines can also be used to collect hashes of popular misleading or hateful content. Hashes are small 'signatures' or 'fingerprints' that do not contain the original content but can be used to identify very similar content. Hashes can thus be used to develop on-device solutions that work in encrypted settings. For instance, Reis et al. (2020) examine images and propose an on-device approach to alerting users to content that has been fact-checked on WhatsApp. Their solution focuses on PDQ hashes for images and requires a list of hashes for known pieces of misinformation. Our analysis in this paper suggests that tiplines could be a successful way to populate such a list. The most popular images are likely to be submitted to a tipline, and, even better, they are very likely to be submitted to the tipline before they are widely shared within public groups. Thus, if a list was populated based on images sent to tiplines, it might identify many these shares.

Using advances in the state-of-the-art techniques to find similar image and text messages, an on-device fact-checking solution could identify up to 40% of the shares of potential misinformation in public WhatsApp groups while preserving end-to-end encryption if content can be prioritized appropriately and responded to quickly. Such a solution could operate similar to personal antivirus software where individuals can choose from a variety of vendors and fully control what happens when a potential match is identified.

Findings

Finding 1: Tiplines capture content quickly; popular content often appears in tiplines before appearing in public groups.

We examined the effectiveness of tiplines in three ways: speed (i.e., how quickly new content appears in tiplines), overlap with content in public groups, and volume. We began with speed and first examined how long it took for an item to be shared by someone to the tipline. The intuition behind this is that one facet of an effective solution is its ability to identify potential misleading content quickly before it spreads widely.

Figure 1 shows the time difference between an image being shared on a public group and the tipline. Negative values on the x-axis indicate that the content was shared in a public group first. We see that roughly 50% of all the content was shared in public groups first, with around 10% of content going back to over a month. However, if we focus on the subset of the top-10% most shared images within the public groups, the distribution looks very different. We clearly see that a majority of the content (around 80%) was shared on the tipline before being shared in the public groups, indicating that the tipline does a good job covering the most-shared content quickly. Similar trends exist for images on ShareChat (Figure 2). In fact, images sent to the tipline have significantly more shares (41 vs. 29) and likes (51 vs. 40) on ShareChat compared to images not sent to the tipline ($p < 0.01$ for a t-test of means).

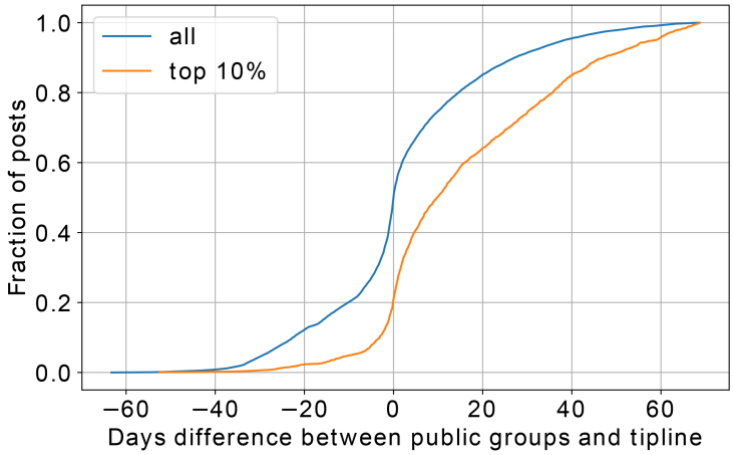

Figure 1. Time difference between the sharing of images on public groups and the tipline. *Approximately 50% of the images were shared on public groups first. However, if we consider just the top 10% most shared images in the public groups, they were mostly shared first on the tipline. (Negative values on the x-axis represent items being shared in the public groups before being shared on the tipline.)*

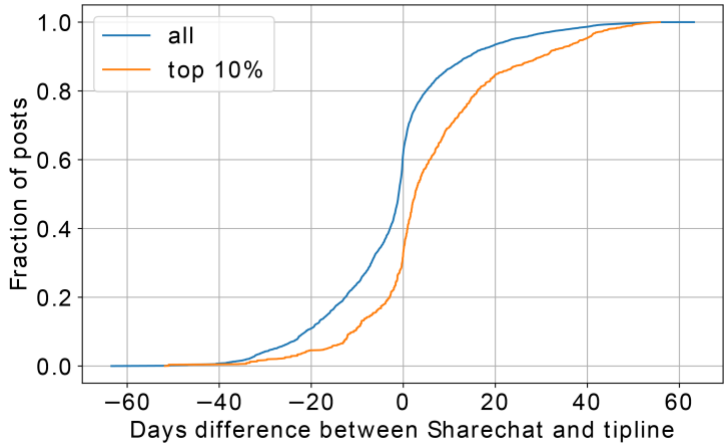

Figure 2. Time difference for images shared on ShareChat and the tipline*. The most popular content was more likely to be shared on the tipline first compared to all content.*

Comparing the text messages within the public groups to the tipline messages leads to similar results (Figure 3). To make this comparison, we first clustered all text messages in the public groups and, separately, in the tipline. This comparison only uses the text messages from the tipline within clusters having at least five unique messages that were annotated as having claims that could be fact-checked to avoid the risk of matching spam or less meaningful content. We again find that the most shared content was often shared to the tipline before spreading widely within the public groups. Similar trends also exist for URLs (Figure 3, green and red lines).

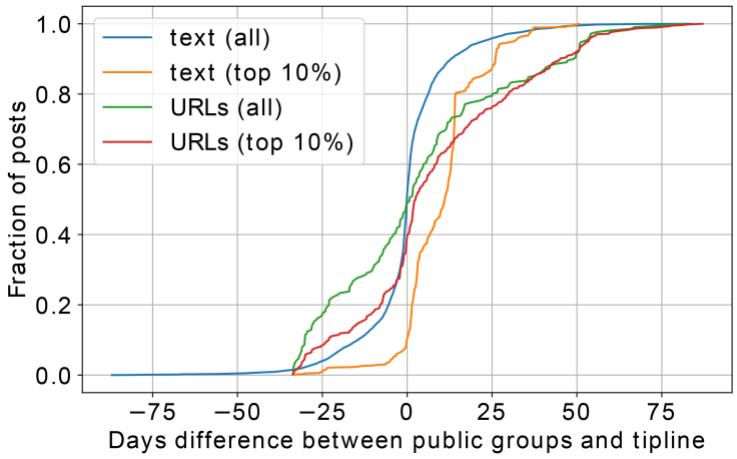

Figure 3. Time difference between the sharing of text messages and URLs in the WhatsApp tipline and public groups.

These findings suggest that content submitted to the tipline may have been circulating person-to-person or in smaller, private groups not in our data before the content was submitted to the tipline or appeared in the large, public groups in our data. Popular content on non-encrypted social media platforms often spreads quickly through large broadcast events (Bright, 2016; Goel et al., 2015); such broadcast events may be rarer on WhatsApp, however, due to the limits on message forwarding and the size of groups.[8]

Finding 2: Tiplines capture a meaningful percentage of content shared in public groups.

A second facet of effectiveness is content overlap: for tiplines to be an effective source of content for fact-checking, we would want them to identify content spreading in other sources of data, including WhatsApp public groups, fact checks, and open social media platforms. We first examined the coverage and computed the number of shares for images in the public groups or on ShareChat and computed what percentage of the images with different numbers of shares appear in the tipline dataset. Figures 4 & 5 show the results. For both the public groups and ShareChat, we used logarithmic bucketing of the number of shares of items to estimate message popularity. The results show tiplines have good coverage of popular content: 67% of the images shared more than 100 times in the public groups were also submitted to the tipline. We repeated the analysis with text messages and found that 23% of text messages shared more than 100 times in the public groups were also submitted to the tipline (Figure 6). To put matters into perspective, we conducted a similar experiment matching all the fact-checked text claims and their corresponding social media posts from the same time period against WhatsApp public groups messages. Only 10% (12/119) of textual content from popular clusters in public groups (shared more than 100 times) matched with at least one text (claim or fact-checked tweet) from Indian fact-checks during this period.

[8] At the time of writing, WhatsApp limits groups to a maximum of 256 people and allows messages to be forwarded to a maximum of five groups at once. If a message has been forwarded at least five times, it can only be forwarded to one additional group at a time. Further details are at https://faq.whatsapp.com/kaios/chats/how-to-create-a-group/?lang=en and https://faq.whatsapp.com/general/chats/about-forwarding-limits/?lang=en

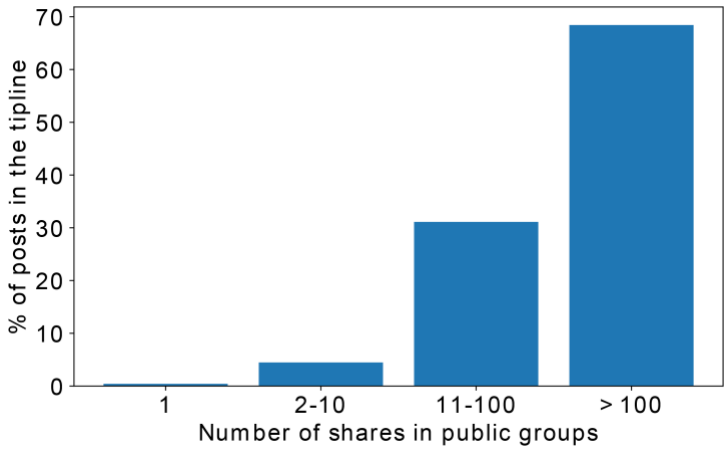

Figure 4. Coverage of images. *The x-axis shows the number of shares on the public groups and the y-axis shows the percentage of images with x shares that match with an image submitted to the tipline. Images that are highly shared on the public groups are much more likely to be shared to the tipline.*

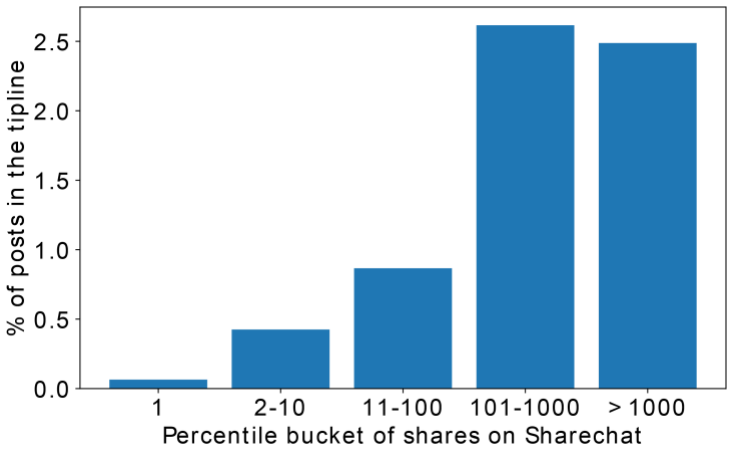

Figure 5. Coverage. *Similar to Figure 4, images shared more often on ShareChat are more likely to appear in the tipline.*

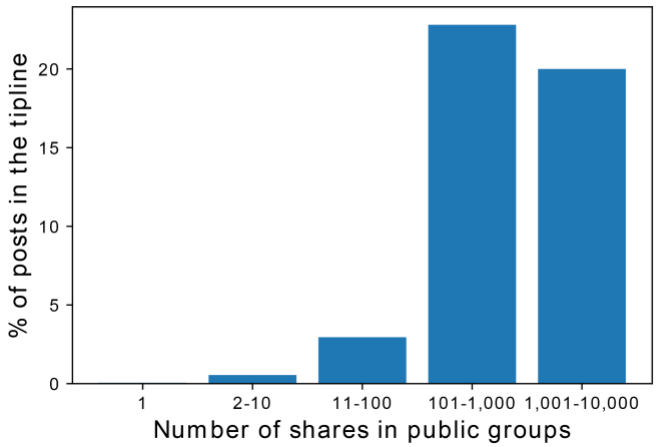

Figure 6. Coverage of text messages. *The x-axis shows the number of shares on the public groups and the y-axis shows the percentage of text messages with x shares that match with a text message submitted to the tipline. Text messages that are highly shared on the public groups are much more likely to be shared to the tipline. Messages in the public groups are first clustered together to determine the number of shares of each message.*

Exact copies of about 10% of popular URLs (i.e., URLs shared over 1,000 times) on public groups were also submitted to the tipline. Because of shortened URLs, content takedowns, and the 2-year time difference between data collection and analysis, grouping URLs was very challenging. We therefore limited further analysis of URLs for this research question.

We found many text messages and images submitted to the tipline did not appear in the public groups, which suggests tiplines also capture content being distributed in WhatsApp in smaller-group or person-to-person settings. Out of the 23,597 unique clusters of images submitted to the tipline, only 5,811 clusters (25%) had at least one match with an image from the public groups.

Next, we checked which text messages from the clusters with claims matched messages found in the public group data. We found that 93% of the 257 relevant clusters match at least one message in the WhatsApp public group dataset. Far from being a skewed result where only few large clusters match, we found a large number of messages across clusters of all sizes match at least one public group message. The per-cluster average of tipline messages matching to the public group data is 91%. This suggests that if we had included clusters with fewer than five unique messages, we may have seen additional matches. We did not include these, as we only wanted to include messages we knew had fact-checkable claims (and we only annotated clusters with at least five unique messages). Additional annotation would likely yield more relevant messages and matches.

Seven percent of the text clusters with fact-checkable claims from the tipline did not match any public group messages. This implies that collecting messages from public groups and using tiplines can be complementary even though neither is a full sample of what is circulating on WhatsApp.

Finally, we measured the potential impact tiplines could have on preventing the spread of misinformation. For this, we looked at items that were shared on both the tipline and in the public groups. We identified the timestamp when an item was first shared on the tipline and counted the number of shares of the item on the public groups before and after this timestamp. The intuition here is that if an item was shared on the tipline, it is in the pipeline to be fact-checked. We found that 38.9% of the image shares and 32% of the text message shares in public groups were after the items were submitted to the tipline.

Finding 3: Tiplines capture diverse content, and a large percentage of this content contains claims that can be fact-checked.

To investigate the third research question, we took an in-depth look into images, text messages, and links sent to the tipline, and here we present examples of the most popular submissions.

Images

The tipline received 34,655 unique images, which clustered into 23,597 groups. Figure 7 shows the three most submitted images to the tipline. Each of these three images was submitted by at least 60 unique users. All three of these images were fact-checked and found to be false. Figure 7a shows a 'leaked' government circular alleging a terrorist plot during the elections. This was in fact an old circular taken out of context. Figure 7b falsely alleges that Pakistani flags were raised during a political rally, and Figure 7c shows doctored screenshots of a TV news program.

(a) (b) (c)

Figure 7. The most shared images on the tipline.

We constructed a visual summary of all the unique images sent to the tipline, as shown in Figure 8. The mosaic shows various categories of images sent to the tipline at a high level. As we move from the top left to the bottom right, we can see a lot of images on the top left of Figure 8 containing pictures of newspapers, and in general images with text. As we go to the bottom left, we see memes and pictures containing quotes of politicians, and on the bottom right, images of people/politicians. Pictures of newspapers or images with text on them are the most dominant type, constituting over 40% of the content, followed by memes which make up roughly 35% of the content.

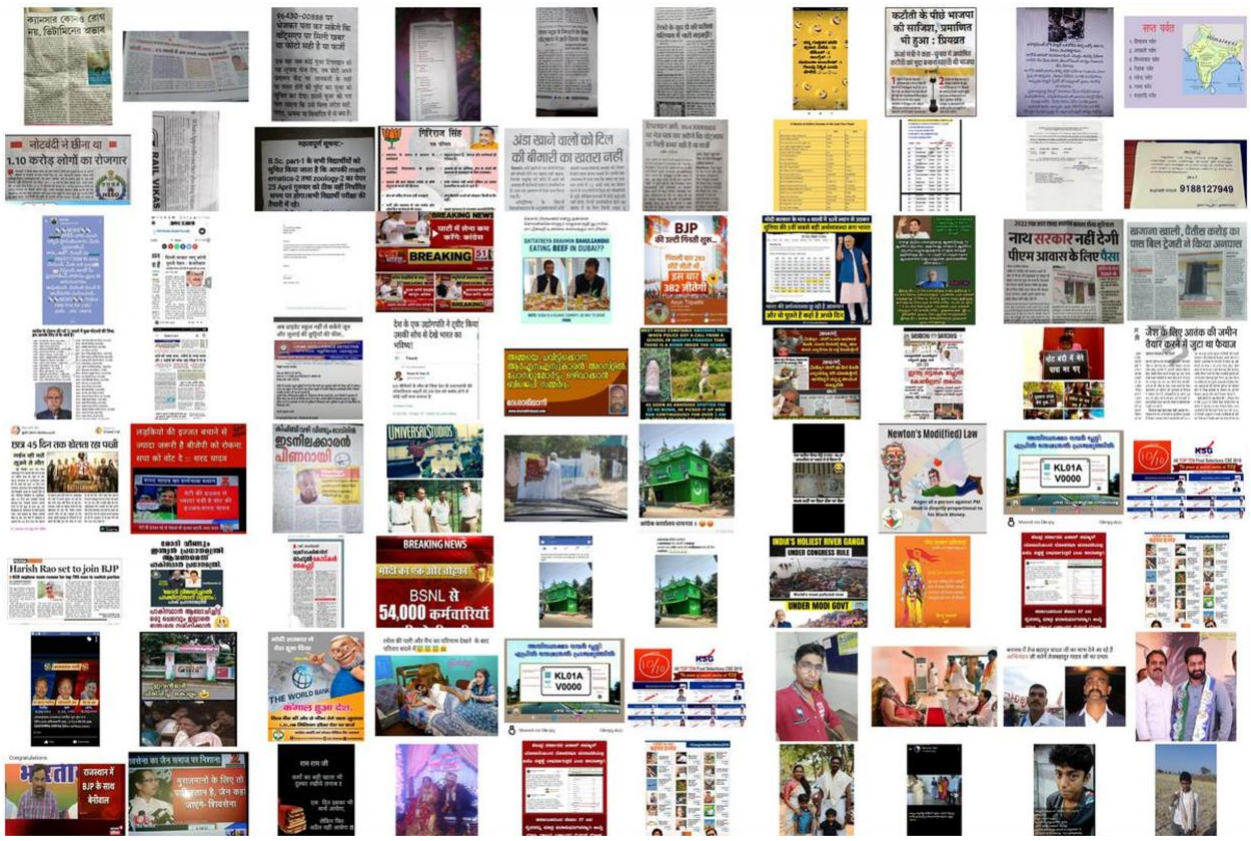

Figure 8. A visual summary of the images submitted to the tipline. *The mosaic is a collection of 20 clusters obtained from the 34k images submitted to the tipline. Each cluster is represented as 2x2 grid of images randomly sampled from the cluster.*

Text messages

Of the 88,662 text messages sent to the tipline, 37,823 are unique (not exact duplicates). We further organized the messages by clustering them using the Indian XLM-R model (Kazemi et al., 2021) and a threshold of 0.9, which resulted in 20,856 clusters (or groups) of near duplicate messages. Each cluster represents a group of text messages with nearly the same meaning. There were 559 clusters with five or more unique messages. We hired an Indian journalist with fact-checking experience during the 2019 Indian general election to annotate each of these clusters for the quality of the clustering and to identify clusters with claims that could be fact-checked as defined by Konstantinovskiy et al. (2021), which excludes several statement categories such as personal experience and spam. The annotation interface presented three examples from each cluster: one with the lowest average distance from all other messages in the cluster, one with the highest average distance from all other messages in the cluster, and one message chosen randomly. We found 257 clusters (out of the 559, 46%) comprising 2,536 unique messages were claims that could be fact-checked. Overall, 173 clusters (1,945 unique messages, 7,131 total messages) were related to the election, and 84 clusters (591 unique messages, 2,473 total messages) were claims unrelated to the election.

The clusters were generally all high-quality: in 98% of the clusters all three messages made the same claim. In 2% of the clusters (11 clusters, 159 unique messages) the three items annotated should not have been clustered together.

There were also 231 clusters that did not have fact-checkable claims. These were usually advertising/spam (114 clusters, 1,245 unique messages) or messages specific to the tipline (177 clusters, 2,957 unique messages). The tipline-specific messages include messages following up on submitted pieces of content, requests for more information about the tipline, and requests for fact checks in additional languages.

We took the 257 clusters that were annotated as containing claims and found that 203 contained messages in only one language (usually Hindi) while the other clusters contained between two and six languages. Languages were detected via CLD3 and were selected when a known language was detected and that detection was reported as reliable by CLD3.[9]

Within the clusters with election-related claims, the largest cluster was misinformation advising voters to ask for a "challenge vote" or "tender vote" if they find they are either not on the voter list or have been marked as already voting.[10] There were 213 unique messages totaling 2,121 submissions to the tipline with this claim across five languages. Other prominent themes within the election-related clusters included messages attacking BJP leader Narendra Modi, pro-BJP messages, and messages criticizing Indian National Congress Party leader Rahul Gandhi.

The largest cluster with a non-election claim was misinformation about the tick marks on WhatsApp. It claims that three blue tick marks indicate the government had observed the message.[11] There were two clusters with different variants of this claim totaling 78 unique messages and 1,000 submissions across Malayalam and English.

Of the 2,536 messages in the clusters containing claims, Hindi (47%), English (35%), and Malayalam (6%) were the most common languages. Marathi, Telugu, and Tamil each accounted for roughly 2% of the messages. This likely reflects both the socio-linguistic characteristics of India as well as the fact that the tipline was most heavily advertised in Hindi and English.

[9] https://github.com/google/cld3
[10] https://archive.is/BWsqR
[11] https://archive.is/WfeRe

In total, there were 9,604 submissions to the tipline comprised of 2,536 unique messages annotated as containing fact-checkable claims (i.e., 7,068 submissions within the set are exact duplicates). It took an average of 5 hours (*SD* = 1.4) for half of the total number of submissions in each of the clusters with claims to arrive to the tipline. 90% of the submissions in each of these clusters arrived within an average of 128 hours (*SD* = 17). This suggests slightly slower dynamics than those that have been seen with the signing of petitions (Margetts et al., 2015) and the sharing of news stories on non-encrypted social media (Bright, 2016).

URLs

Another common content type in WhatsApp groups and tiplines is URLs. The tipline received 28,370 URLs (12,674 unique URLs), which contained URLs from 2,781 unique domains. A list of most frequent domains is presented in Table 2. The most prevalent websites submitted to the tipline were social media (YouTube, Facebook, Twitter, and Blogger), news outlets (IndiaTimes and DailyHunt), and URL shortening services (Bitly and TinyURL).

Table 2. *Top 10 domains most shared on the WhatsApp tipline around the Indian general election.*

Domain	Total URLs
YouTube	2,350
Blogger	2,107
Bitly	1,636
Google	1,471
Facebook	1,192
RechargeLoot	724
IndiaTimes	587
DailyHunt	574
Twitter	515
TinyURL	465

Methods

Image similarity. To identify similar images, we used Facebook's PDQ hashing algorithm and Hamming distance. PDQ is a perceptual hashing algorithm that produces a 64-bit binary hash for any image. Small changes to images result in only small changes to the hashes and thus allow visually similar images to be grouped. This allows, for instance, the same image saved in different file formats to be identified. For this paper, images with a Hamming distance of less than 31 were considered to be similar. The same threshold was used previously by Reis et al. (2020). Similar images were clustered together using the DBSCAN (Ester et al., 1996) algorithm.

To construct the visual summary of the images shown in Figure 8, we first obtained a 1,000-dimensional embedding for each image using a pretrained ResNeXt model (Xie et al., 2017). Next, we clustered these embeddings using a k-means clustering algorithm and chose k = 20 using the elbow method. For each cluster, we picked four randomly sampled images and created a mosaic of the 20 clusters.

Text similarity. To identify similar textual items, we used a multilingual sentence embedding model trained for English, Hindi, Bengali, Marathi, Malayalam, Tamil, and Telugu (Kazemi et al., 2021). Kazemi et al. (2021) evaluated this model for claim matching using similar data and found applying a cosine similarity threshold of 0.9 to pairs of messages resulted in the best performance, with an overall F1 score of 0.73. The model performs better on English and Hindi (which are 82% of our data), with an average F1 score of 0.85. Throughout this paper we used a cosine similarity threshold of 0.9 for matching text items.

Text clustering. We clustered text items using online, single-link hierarchical clustering. Each new message arriving to the tipline was compared to all previous messages, and the best match found. If this match was above the similarity threshold, then we added the new message to the same cluster as the existing message. We applied the same process to the public group messages. To enable quick retrieval, we constructed a FAISS (Johnson et al., 2017) index using our Indian XLM-R embeddings of all the public group messages. We then queried this index for each tipline message and recorded all matches with a cosine similarity score of at least 0.9. We remove any duplicate matches (i.e., cases where two tipline messages matched the same public group message) before analyzing the matches.

Data

We used a wide range of data sources in this work including WhatsApp tipline data, social media data from WhatsApp public groups and ShareChat, and published fact checks. All the data used pertains to the four-month period between March 1, 2019, and June 30, 2019. This period includes the 2019 Indian general election, which took place over a period of six weeks in April and May 2019.

Tiplines. In 2019, PROTO led the Checkpoint project using Meedan's open-source software to operate a WhatsApp tipline. PROTO advertised their WhatsApp number asking users to forward any potentially misleading content related to the election. They advised that they would be able to check and reply to some of the content that they received. Over the course of four months, 157,995 messages were received. Of these, 82,676 were unique and consisted of 37,823 text messages, 10,198 links, and 34,655 images. We obtained a list of links, text messages, and images along with the timestamps of when they were submitted to the tipline. We have no information about the submitting users beyond anonymous ids.

WhatsApp public groups. There are currently over 400 million active WhatsApp users in India. With the availability of cheap Internet data and smartphones with WhatsApp pre-installed, the app has become ubiquitous. Aside from messaging friends and family, Indians use WhatsApp to participate in political discourse (Farooq, 2017). Political parties have taken this opportunity to create thousands of public groups to promote their political agendas. These groups have been shown to be quite prevalent, with over one in six Indian WhatsApp users belonging to at least one such group (Lokniti, 2018).

In addition to the image and text items submitted to the tipline, we have data from large "public" WhatsApp groups collected by Garimella and Eckles (2020) during the same time period as the tipline ran. The dataset was collected by monitoring over 5,000 public WhatsApp groups discussing politics in India. For more information on the dataset, please refer to Garimella and Eckles (2020).

ShareChat. ShareChat is an Indian social network that is used by over 100 million users.[12] It has features similar to Instagram and is primarily multimedia focused (Agarwal et al., 2020). Unlike WhatsApp, ShareChat provides global popularity metrics including likes and share count, which allowed us to

[12] https://sharechat.com/

construct a proxy for the popularity of the content on social media. ShareChat curates popular hashtags based on topics such as politics, entertainment, sports, etc. During the three months of data collection, every day, we obtained the popular hashtags related to politics and obtained all the posts containing those hashtags. This provides a large sample of images related to politics that were posted on ShareChat during the data collection period (March 1 to June 30, 2019).

Fact checks. We also collected fact checks and social media data from the time period in English and Hindi. We crawled popular fact-checking websites in India and obtained articles and any tweets linked within the articles following the approach of Shahi (2020) and Shahi et al. (2021). Overall, we found 18,174 fact-check articles in 49 languages from 136 fact checkers from all over the globe. To select fact checks concerning the Indian general election, we filtered the data to require that either the fact check be written in an Indian language or the fact-checking domain be within India's country code top-level domain.

In total, we obtained 3,224 and 2,220 fact checks in English and Hindi respectively. The fact checks were of content from various social media platforms, including Twitter. Whenever available, we obtained the links to the original tweets that were fact-checked and downloaded these. We obtained 811 tweets in total, 653 (182 unique) in English and 158 (63 unique) in Hindi.

A summary of all the data collected is shown in Table 3.

Table 3. *Datasets used in this work. The values shown in parentheses indicate the number of unique messages/images. We only collected image data from ShareChat.*

Dataset	# Text messages (unique)	# Images (unique)
Public groups	668,829 (445,767)	1.3M (977,246)
ShareChat	-	1.2M (401,137)
Checkpoint	88,662 (37,823)	48,978 (34,655)
Fact-check articles	5,444 (5,444)	-
Fact-check tweets	811 (245)	-

Note: M denotes million.

Bibliography

Agarwal, P., Garimella, K., Joglekar, S., Sastry, N., & Tyson, G. (2020). Characterising user content on a multi-lingual social network. *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020), 14*, 2–11. https://ojs.aaai.org/index.php/ICWSM/article/view/7274

Arun, C. (2019). On WhatsApp, rumours, and lynchings. *Economic & Political Weekly, 54*(6), 30–35. https://www.epw.in/journal/2019/6/insight/whatsapp-rumours-and-lynchings.html

Bright, J. (2016). The social news gap: How news reading and news sharing diverge. *Journal of Communication*, *66*(3), 343–365. https://doi.org/10.1111/jcom.12232

Elkind, P., Gillum, J. & Silverman, C. (2021). *How Facebook undermines privacy protections for its 2 billion WhatsApp users*. The Wire. https://thewire.in/tech/facebook-undermines-privacy-protections-whatsapp-users

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 14th International Conference on Data Engineering, 96,* 226–231. https://doi.org/10.1109/icde.1998.655795

Farooq, G. (2017). Politics of fake news: How WhatsApp became a potent propaganda tool in India. *Media Watch, 9*(1), 106–117. https://doi.org/10.15655/mw/2018/v9i1/49279

Garimella, K., & Eckles, D. (2020). Images and misinformation in political groups: Evidence from WhatsApp in India. *Harvard Kennedy School (HKS) Misinformation Review, 1*(5). https://doi.org/10.37016/mr-2020-030

Goel, S., Anderson, A., Hofman, J., & Watts, D. (2015). The structural virality of online diffusion. *Management Science, 62*(1), 180–196. https://doi.org/10.1287/mnsc.2015.2158

Hassan, N., Li, Chengkai, & Tremayne, M. (2015). Detecting check-worthy factual claims in presidential debates. *CIMK '15: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (pp. 1835–1838). Association for Computing Machinery. https://doi.org/10.1145/2806416.2806652

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, *7*(3), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

Kazemi, A., Garimella, K., Gaffney, D., & Hale, S. A. (2021). Claim matching beyond English to scale global fact-checking. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* (pp. 4504-4517). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.347

Konstantinovskiy, L., Price, O., Babakar, M., & Zubiaga, A. (2021). Towards automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital Threats: Research and Practice*, *2*(2). https://doi.org/10.1145/3412869

Lo, K. (2020, December 14). *Fact-checking and mental health*. Meedan. https://meedan.com/blog/fact-checking-and-mental-health/

Lokniti, C. (2018). *How widespread is WhatsApp's usage in India?* Mint. https://livemint.com/Technology/O6DLmIibCCV5luEG9XuJWL/How-widespread-is-WhatsApps-usage-in-India.html

Lomas, N. (2019). *WhatsApp adds a tip-line for gathering fakes ahead of India's elections*. TechCruch. https://techcrunch.com/2019/04/02/whatsapp-adds-a-tip-line-for-checking-fakes-in-india-ahead-of-elections/

Maldita.es (2021). *Disinformation on WhatsApp: Maldita.es' chatbot and the "frequently forwarded" attribute.* https://web.archive.org/web/20211129201556/https:/maldita.es/uploads/public/docs/disinformation_on_whatsapp_ff.pdf

Margetts, H., John, P., Hale, S., & Yasseri, T. (2015). *Political turbulence: How social media shape collective action.* Princeton University Press. https://doi.org/10.2307/j.ctvc773c7

Meedan. (2020, December 7). *One year of running the WhatsApp end-to-end fact-checking project.* https://meedan.com/blog/one-of-year-of-running-the-end-end-to-fact-checking-project/

Meedan (2021). *FACT CHAMP: New project to increase collaboration between fact-checkers, academics, and community leaders to counter misinformation online*. https://meedan.com/blog/fact-champ-launch/

Melo, P., Messias, J., Resende, G., Garimella, K., Almeida, J., & Benevenuto, F. (2019). WhatsApp monitor: A fact-checking system for WhatsApp. *Proceedings of the International AAAI Conference on Web and Social Media, 13*(1), 676–677. https://ojs.aaai.org/index.php/ICWSM/article/view/3271

Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., Papotti, P., Shaar, S., & Da San Martino, G. (2021). Automated fact-checking for assisting human fact-checkers. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21* (pp. 4551–4558). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2021/619

Reis, J., Melo, P., Garimella, K., & Benevenuto, F. (2020). Can WhatsApp benefit from debunked fact-checked stories to reduce misinformation? *Harvard Kennedy School (HKS) Misinformation Review, 1*(5). https://doi.org/10.37016/mr-2020-035

Resende, G., Melo, P., Sousa, H., Messias, J., Vasconcelos, M., Almeida, J., and Benevenuto, F. (2019). (Mis)information dissemination in WhatsApp: Gathering, analyzing and countermeasures. *WWW '19: The World Wide Web Conference* (pp. 818–828). Association for Computing Machinery. https://doi.org/10.1145/3308558.3313688

Shaar, S., Babulkov, N., Da San Martino, G., and Nakov, P. (2020). That is a known lie: Detecting previously fact-checked claims. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 3607–3618). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.acl-main.332

Shaar, S., Hasanain, M., Hamdan, B., Ali, Z.S., Haouari, F., Nikolov, A., Kutlu, M., Kartal, Y.S., Alam, F., Da San Martino, G., Barrón-Cedeño, A., Míguez, R., Beltrán, J., Elsayed, T., & Nakov, P., (2021). Overview of the CLEF-2021 CheckThat! Lab Task 1 on check-worthiness estimation in tweets and political debates. In G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), *Proceedings of the Working Notes of CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania*. http://ceur-ws.org/Vol-2936/paper-28.pdf

Shahi, G. K. (2020). *Amused: An annotation framework of multi-modal social media data*. ArXiv. https://arxiv.org/abs/2010.00502.

Shahi, G. K., Dirkson, A., & Majchrzak, T. A. (2021). An exploratory study of COVID-19 misinformation on Twitter. *Online Social Networks and Media*, *20*. https://doi.org/10.1016/j.osnem.2020.100104

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1492–1500). IEEE. https://doi.org/10.1109/cvpr.2017.634

Acknowledgements
We are grateful to Meedan colleagues, Pop-Up Newsroom, PROTO, and Prof. Rada Mihalcea for valuable feedback and data access.

Funding
This work was funded by the Omidyar Network with additional support from Sida, the Robert Wood Johnson Foundation, and the Volkswagen Foundation. Kiran Garimella was supported by the Michael Hammer postdoctoral fellowship at MIT.

Competing interests
Meedan is a technology non-profit that develops open-source software that fact-checking organizations use to operate misinformation tiplines on WhatsApp and other platforms. The research team at Meedan operates independently: the research questions, approach, and methods used in this article were decided by the authors alone.

Ethics
Throughout this research, we were extremely concerned about the privacy of tipline users and the ethical concerns that come with large-scale studies. All WhatsApp messages in our datasets were anonymized, and personal information (e.g., phone numbers) was removed. Our experiments were done at the macro level, and we followed strict data access, storage, and auditing procedures to ensure accountability.

Data availability
For privacy reasons, the raw content of the WhatsApp messages used in this study cannot be released. Metadata from the WhatsApp tipline and public groups are available at https://doi.org/10.7910/DVN/ZQWG02. ShareChat data is available from Agarwal et al. (2020) at https://nms.kcl.ac.uk/netsys/datasets/share-chat/