Journal of Quantitative Description: Digital Media 1(2021), 1â n 2673-8813/2021.XXXXX

What Circulates on Partisan *WhatsApp* in India? Insights from an Unusual Dataset.

SIMON CHAUCHARD

University Carlos III Madrid, Spain

KIRAN GARIMELLA

Rutgers University, USA

WhatsApp is the most used app on Android, with over 2 billion monthly active users. With this popularity, political actors in countries ranging from the Philippines to Brazil have embraced WhatsApp as a medium to communicate with voters. In India, WhatsApp groups backed by political parties are suspected of spreading misinformation and/or of circulating hateful content pointed towards minority groups, potentially leading to offline violence. They are also often described as a powerful propaganda tool for the ruling party (the BJP). Yet, despite these narratives, we so far know little about the content that circulates on these partisan groups. In this manuscript, we describe the visual content of 533 private WhatsApp groups maintained by party workers across the state of Uttar Pradesh, collected over a period of 9 months. Manual coding of around 40,000 images allows us to estimate the amount of misinformation/hateful content on one hand, and partisan content on the other. Additional matching of this data with other sources in turn allows us to evaluate the extent to which the content posted on these local WhatsApp threads corresponds to the content posted by the party leadership. Analyses suggest that local partisan threads contain few hateful or misinformed posts; more surprisingly maybe, most content cannot easily be classified as "partisan". While much content appears to be religion-related, which may serve an indirect partisan role, the largest share of the content is neither partisan nor religious, and more easily classifiable as phatic (Berriche and Altay, 2020) or entertainment related.

Simon Chauchard: simon.chauchard@uc3m.es
Kiran Garimella: kg766@comminfo.rutgers.edu
Date submitted: 2022-01-16

Keywords: WhatsApp; Social Media; India; Misinformation; Hateful Content; Political Communication

Introduction

Encrypted messenging apps such as WhatsApp allow for rapid and private communication within and across communities of users organized as groups. With over 2 billion monthly active users, WhatsApp is popular all around the world. In developing countries such as India and Brazil, users rely on WhatsApp in a variety of ways, including as a social network to stay in touch with their close connections, and to consume news and other information. With the growing popularity of WhatsApp, political parties have started creating specific WhatsApp communication strategies, including adding normal users to larger, less intimate groups in which they receive social or political news from more distant connections (Lupu et al., 2020; Evangelista and Bruno, 2019; Resende et al., 2019; Farooq, 2017). Such groups organized by partisan actors are reportedly used by over one in six WhatsApp users in India and Brazil (Lokniti, 2018; Newman et al., 2019). WhatsApp allows groups with up to 256 members, where political parties share multimedia messages, transforming presumably innocuous chat groups into highly active social spaces for the dissemination of information. The fact that posts on chat apps are private and protected by encryption means that no-one, including WhatsApp itself, gets to see, read, filter or analyze content, and that it is impossible to trace the source or extent of spread of a message on the platform.

In recent years, threads associated to political actors have been suspected of conveying misinformation and/or hateful content pointed towards minority groups, potentially leading to offline violence (Perrigo, 2019; Cheeseman et al., 2020; Zizumbo-Colunga and del Pilar Fuerte-Celis, 2020; Nuraniyah, 2019). In India, the focus of this study, recent news reports for instance suggest that WhatsApp groups of the ruling Bharatiya Janta Party (BJP) "demonise(d) the Muslim community through unverified statistics and 'data', call for violence against them and celebrate the prospect of Muslims losing citizenship" (Purohit, 2019). These threads are also often described as an efficient propaganda tool for the ruling party (the BJP): partisan WhatsApp groups have been credited with the repeated success of the BJP (Schakel et al., 2019). In both cases, this dissemination has been hypothesized to take place in a top-down manner, with higher-level actors within the party (the famed "IT cell") encouraging the coordinated dissemination of partisan materials among lower-level actors (admins of the local-level groups). According to this common narrative, the BJP's WhatsApp groups have since the early 2010s allowed the party to spread party messages in a direct, unfiltered and efficient manner, creating

legions of loyal party followers throughout India.[1]

Yet, despite much suspicion, we so far have little quantitative evidence on the actual content of these partisan groups. In this project, we rely on unusual and so far unique data collected in conjunction with a survey of Indian parties' booth-level "social media workers" (the aforementioned lower-level actors in charge of maintaining WhatsApp groups), in order to document the content of such partisan WhatsApp groups. Specifically, we describe the visual content of a sample of 533 closed groups from multiple political parties whose "admin" was a party worker from across the state of Uttar Pradesh (the most populated Indian state with over 220 million people, where the BJP is currently is in power).[2] The content of these threads was exported over a period of 9 months; A fine grained coding of a random sample of this content (around 40,000 images) by human coders allows us to evaluate the amount and the type of misinformation/hateful content on one hand, and the amount of content classifiable as "partisan" on the other.[3] Beyond partisan posts, we provide a classification of the residual visual content of these threads — i.e., the content that does *not* easily fit within both or either of these categories. Altogether, we provide a credible description of the visual content posted and/or forwarded on these partisan threads.

While groups affiliated to the ruling party (the BJP) constitute the bulk of our sample, we are able to compare the behavior of posters on these threads to the behavior of posters on other parties' threads. Several other types of comparisons allow us to refine our inferences. We first compare the posting behavior of group "admins" to the behavior of simple users. Second, in order to document potential top-down transmission of content, we compare the content posted on BJP threads in our sample to the content posted by party officials on the very

[1]Similar dynamics are said to exist elsewhere. In Brazil, reports for instance suggest that Jair Bolsonaro's election prospects were enhanced by the activism of a structured pro-Bolsonaro WhatsApp propaganda machine, which successfully managed to "dismiss the political system and mainstream media as corrupt, as well as targeted leftwing politicians and activists, often using homophobic tropes and anti-feminist slurs" (Avelar, 2019). See also Reis et al. (2020) on this.

[2]More details on sampling in Section 2.

[3]Note that these categories are *not* mutually exclusive.

official and institutional *NaMo* app[4]. This allows us to evaluate whether the content posted on local threads differs from the content promoted by high-level party actors.

Results suggest that misinformation and hateful content, while they account for an important part of some sub-types of content (for instance, content about minorities on the ruling party's threads), remain overall rare: they account for only a few percents of the *total* content posted on these partisan threads, including on BJP threads. More surprisingly maybe, *most* of the content cannot easily be classified as "partisan content". Salutations and wishes, often formulated in religious terms and/or relying on religious iconography, constitute much of the content. A large share of the content is also neither partisan nor religious, and more easily classifiable as phatic (Berriche and Altay, 2020) or entertainment-related. Only a small share (19%) of NaMo app content eventually ends up on the WhatsApp threads. More importantly, it is far from being the content most frequently shared by users on the threads, which suggests – coherent with our codings – that this content would most likely be drawn in a sea of other, unrelated content. Altogether, our results thus suggest that *potential* persuasion or mobilization effects of partisan WhatsApp threads happen in a context in which most of the content posted *at best* plays an indirect role in achieving these effects. In our view, this should crucially inform theoretical arguments developed about the effect these threads might have, insofar as powerful effects may derive less from the frequency of polarizing content than to its relative rarity or to its combination with other types of content.

These findings expand our understanding of social media diets and of the mechanisms through which misinformation operates to new contexts and new media. Since the 2016 US election, social media misinformation has provoked considerable attention in academia. Attempts to study how misinformation is shared (Vosoughi et al., 2018), its prevalence (Allcott and Gentzkow, 2017; Fourney et al., 2017; Grinberg et al., 2019; Guess et al., 2019), why it is believed (Flynn et al., 2017; Pennycook et al., 2018), and its effects on public opinion (Guess et al., 2020a) have led to a rapid and expansive outbreak of scholarly attention on the subject (Lazer et al., 2018; Flynn et al., 2017; Jerit and Zhao, 2020; Wittenberg and Berinsky, 2020). Much of this scholarship has examined the media diets of social media users or the content posted on social media. Without minimizing popular concern about social media misinforma-

[4]https://www.narendramodi.in/downloadapp, which has over 10 million installs on the Android app store.

tion, research on Facebook users' information diet has shown that fewer users than popularly expected were exposed to fake news (Guess et al., 2019); that only specific categories of users shared fake news (Guess et al., 2020b); and that their degree of exposure did not decisively indicate a change in downstream attitudes or behaviors (Guess et al., 2020a).

Due to a mix of ethical and practical difficulties, very little of this scholarship has so far touched on WhatsApp. A handful of recent works have now studied the circulation of information on WhatsApp (Rossini et al., 2020) and discussion dynamics on the platform (Gil de Zúñiga et al., 2019; Valeriani and Vaccari, 2018; Vaccari and Valeriani, 2018); some have even questioned the downstream effects of WhatsApp on participation to social movements (Treré, 2020). But most of these efforts have focused on advanced democracies. Of the few published or forthcoming works on the messaging app (Badrinathan, 2020; Badrinathan and Chauchard, 2021; Garimella and Tyson, 2018; Garimella and Eckles, 2020; Rossini et al., 2020), none have so far studied *private*, closed partisan groups on a large scale.

The rest of the study proceeds as follows: in Section 2, we detail the nature of our dataset. Next, we explain how the dataset was annotated in order to allow us to classify its content, and how we subsequently matched its content to the content of other databases (Section 3). In Section 4, we use these codings to try and quantify the amount of misinformation and hateful content present in our sample of partisan WhatsApp threads. Section 5 explores the extent to which the content of these partisan threads may be construed as partisan content, and if so, what type of partisan content this is. In Section 6, we use our coding to describe the residual (i.e. non-partisan) content that roughly constitute a half of the content on partisan thread. The final Section reflects on our findings.

The Dataset

As noted above, high-quality evidence about Facebook and Twitter users' "information diets" now exists (Guess et al., 2019; Barberá et al., 2015). Generating comparable evidence for chat or discussion apps however requires researchers to overcome several practical and ethical hurdles.

The content of most WhatsApp threads is by design private and encrypted, and as such only accessible to members of the group (that is, members added by one of the "admins").

The processing of data from private WhatsApp groups thus requires researchers to become members of the groups, and/or to obtain its content from members (for instance through a forward). Beyond this technical challenge, the processing and the analyses of WhatsApp threads is also challenging from a legal and ethical point of view: the European Union's current GDPR guidelines, insofar as they typically require consent from every participant to an online discussion, have sometimes been interpreted as suggesting that these analyses should remain off-limit.

In this project, we rely on a creative solution that we also believe to be compatible with current GDPR rules. We rely on the "admins" of partisan WhatsApp threads to get access to the content of the threads.

Sampling

Practically, the research team fielded a survey of political parties' "social media workers" across 10 districts of the Northern Indian state of Uttar Pradesh in March-May 2019, focusing on workers operating at the booth-level (the most local level possible). The districts we collected data for were purposely chosen in collaboration with our implementing partner (the local firm *Across India*). While we made sure to select districts in all major sub-regions of the state so as to ensure a degree of representativeness, our final choice also took into account practical concerns, among which featured the team's experience in that district (which we foresaw to be a likely predictor of our ability to interview respondents). The ten districts are shown (in red) in Figure 1.

For this first step of our data collection, we randomly selected 300 "polling booth areas" (the smallest possible electoral subdivision, counting an average of 1,000 voters) from each of the ten districts highlighted in Figure 1 - aiming for a total across 10 districts of 3,000. Our objective was to interview the social media manager for the ruling party (the BJP) in as many of these as possible. When we could find no social media manager for the BJP at the local level (which could have happened in a number of opposition strongholds), the team interviewed an actor from one of the other parties, likely the locally dominant party.[5] Relying on this strategy,

[5]Either the Samajwadi Party (SP), the Bahujan Samaj Party (BSP) or more rarely, the Indian National Congress (INC).

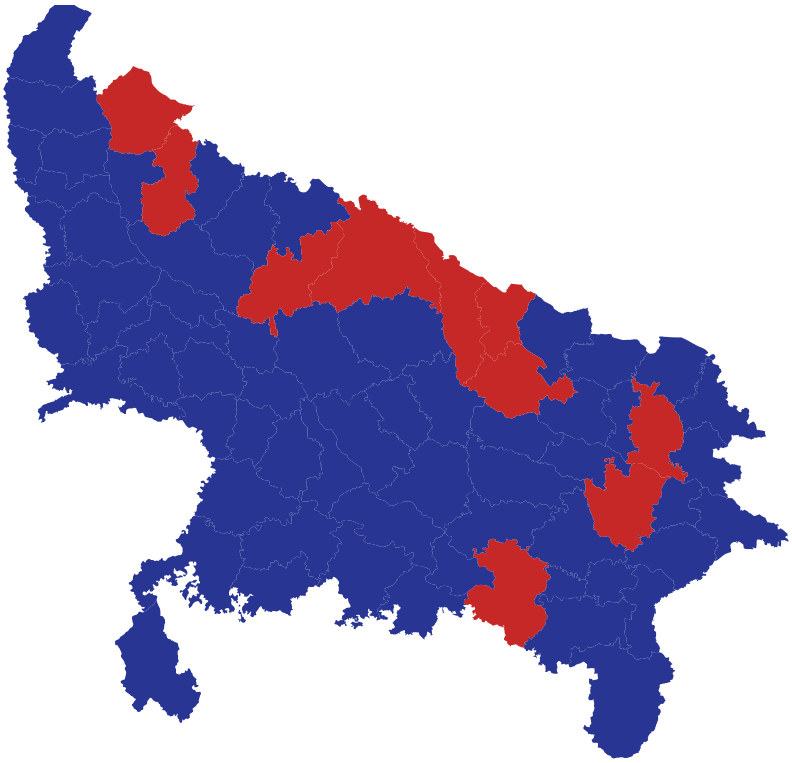

Figure 1. The ten districts in Uttar Pradesh we collected data from in red.

we managed to identify a suitable actor in almost all randomly selected areas: a total of 2,937 social media managers were interviewed, more than 2/3 of which belonged to the BJP. These interviews required these actors to answer *general* questions about social media strategy and party organization.[6]

In a follow-up to these interviews, a *randomly* selected sample of around 50% of the initial sample (n= 1,547) were recontacted and questioned about the *specific* thread(s) they personally maintained.[7] As they were re-interviewed, these actors were asked several inter-related questions. They were first asked to share the name of the *main* partisan group they maintained. 87% of them consented to share that name. If and only if they gave the name of that group, they were then asked to add the research team to the group, so that the research

[6]The results of this survey are analyzed in a related working paper (AUTHOR 2021).

[7]We ended up recontacting closer to 53% of that initial sample as a number of randomly selected actors we initially did not manage to recontact later made contact.

team "can better understand their job and study the content of their threads". If they agreed to this request, the research team gave a phone number for them to add the research team to the thread. 52% of the initial sample of 1,547 (that is, 802 of them) consented to this request during this re-contact. We were however eventually not able to study the threads of each of the social media managers who consented to our request at that stage. This may be due to a variety of reasons. Interviewees may have forgotten to add us or changed their mind after they were given the relevant phone number; they may not have known how to (anecdotally, it appears that some of them did not handle the technical part of the "admin" job themselves); the thread may have been deleted or died by the time we tried to export it (at least a month later); Alternatively, technical difficulties on our end made it difficult to export the content of a small number of the threads.

As a result, overall 34% of the initial sample of 1,547 (533) re-interviewees who had been asked about the main thread they maintained thus added us to a thread whose content we could later exploit. This provides us with the largest to date dataset of such *private* partisan threads. Importantly, we cannot entirely exclude a form of selection bias, since one may suspect that our interviewees *selectively* accepted to let us access these private groups. Several elements however lead us to think that the content we classify is informative and at least somewhat representative of the threads maintained by political parties (especially so the BJP), despite the iterative strategy we relied on to assemble the dataset. It should first be noted that these threads are not necessarily perceived as secretive or even just private by their admins: individuals and numbers are relatively frequently added to these threads on the basis of very weak ties, as party workers face pressure from their hierarchy to add as many numbers as possible to promote the party, in order to be evaluated positively by their hierarchy. By design, it is thus *not* the case that they see these groups are completely shielded from outside monitoring, insofar as they cannot predict what users might themselves do with the data posted; in that sense, the "cost" of adding us to the thread may not have been as steep as if these groups had been properly shielded from observation in the first place. Second, it is unclear to us that these very low-level party affiliates would have had a sense of the type of visual content we might take offense to, or even look forward to count. It seems rather unlikely to us that these actors would themselves see these types of content as shameful or problematic, in a way that would subsequently lead them to withhold access. This is somewhat confirmed by the data we present below, as a majority of threads do contain *some* problematic content (misinformation and/or

hateful content). Additionally, it is important to acknowledge that many of our interviewees would have credibly, *over time*, forgotten the fact that an outsider had been added to the thread. Last but not least, these selection bias concerns would not apply to the content posted by regular users (i.e. "non-admins"), who were never aware of our presence on the thread, and whose posts eventually constitute the majority of the content we analyze. It is thus relatively incredible that our interviewees would have had a sense of the material that was going to be posted *down the line*, since admins do not control the posts of users on WhatsApp, let alone nine months later.

Besides, as shown in Appendix G, we are able to show — relying on data from the initial survey of these actors — that the individuals whose threads are included in our final sample do *not* drastically differ from the individuals included in the initial sample we targeted. While they are somewhat younger, less experienced, and more likely to manage multiple partisan threads, they are equally likely to belong to the BJP (the party we ex-ante would have suspected of being the most strategic and/or suspicious of sharing data with outsiders), including to the most extreme or specialised wings of the party (the aforementioned "IT cell", or the RSS and VHP, two associated organizations known for their ideology, extremism and involvement in hateful acts on the ground). They also appear to engage in similar social media work.

Since the actors who consented to share data were not different on observable variables, we can only speculate as to their motivations for allowing us access to these data. Two inter-related reasons reportedly — based on interviewers' own accounts — explain that a portion of them agreed to our request. The first is that these actors typically were extremely low-level actors in party organizations - and often young men relatively new to the world of politics. Often neglected, or exploited, by their hierarchy, they may have been receptive to an outsider's attention. Second, they may have shared, at least in some cases, some important demographic characteristics with the individuals making the requests (the survey interviewers) on behalf of the research team.

Consent and Ethics

Importantly, the research team did, as part of this process, inform participants of its research intentions and explicitly sought their consent. As we asked them to grant us access to the

partisan WhatsApp thread they maintained, we explicitly stated that "we would like to see the content of these threads for research purposes, to know whether this is indeed a useful tool for political parties." We offered no reward or incentive that might convince them to do otherwise so as not to pressure their choice.

While the research team obtained admins' consent, and dutifully informed them that the research team would protect the identity of *all* participants to these online discussion groups, obtaining formal consent from each and every *user* would have been close to impossible. The reason is simple: these threads can be large (up to 256 participants). This creates a potential ethical hurdle, as the processing of sensitive data without obtaining the consent of all participants (not simply admins) is generally prohibited under GDPR rules.

Our data collection however arguably fulfils one of the conditions that renders the processing of sensitive data compatible with GDPR rules. As per article 6 of the GDPR regulation, the processing of sensitive data obtained *without* the explicit consent of participants is acceptable if it is necessary for research purposes "in the public interest". We expand in Appendix A on the reasons as to why this research arguably qualifies as being "in the public interest." Note, in addition, that this research was reviewed by Leiden University's Data Compliance officer in order to obtain external validation on this criterion. Besides, we refrain from publishing materials in this manuscript that might provide clues about the identity of participants.

What Data Did We Export?

Relying on the aforementioned strategy, the research team was "added" to 533 threads by local party agents. We were only able to monitor and collect data from these groups since the day we were added to a group. Overall we collected 916,157 posts, including 377,265 visual posts (41% of the total), 109,261 videos (11.8% of the total) and 8,392 audio posts. Both the text and visual data from these threads were exported using a *Google chrome* add-on at regular intervals for a period of nine months. The add-on helps us export csv files from the web version of WhatsApp. Personal data contained in the exported data (user phone numbers) were hashed and stored separately in order to preserve the privacy of the users. All subsequent analysis was done on the anonymized data. 340 of the 533 groups (67%) were BJP groups, 48 were SP groups (9%), 12 were BSP and 2 were Congress. The 130 remaining groups were groups

Table 1: Messages per district and party

	BJP	SP	Other	INC/BSP
Azamgarh	89,048	23,544	6,203	0
Bahraich	119,176	1,335	12,165	8,596 (INC)
Bijnor	137,581	3,622	38,530	3,312 (BSP)
Gonda	24,375	4,405	15,219	0
Gorakhpur	46,010	17,158	17,379	12,438 (BSP)
Kheri	159,343	45,709	63,996	3,208 (BSP)
Moradabad	7,675	859	11,006	787 (BSP)
Prayagraj	36,143	4,514	8,469	2,482 (BSP)
Shahjahanpur	31,453	24,665	6,764	0
Shrawasti	67,086	6,078	70,779	0

whose name did not *explicitly* refer to a partisan identity. These included friends groups, village management groups, spam, etc.[8] Table 1 provides key statistics about the distribution of the threads by district and by party. In keeping with the sampling strategy of our survey of social media workers, the dataset covers 10 districts across the state of Uttar Pradesh and mostly contains BJP-related threads.

Additional Data Used in the Study

In some of our analyses, we additionally compare the data from these threads to data drawn from other datasets. We first compare them to data posted through the *NaMo* app.[9] Named after BJP Prime Minister Narendra Modi, *NaMo* is an app designed to allow the social media

[8]Several reasons may explain how these groups ended up in our sample. It may first be the case that the online partisan activity of parties' social media workers takes place on non-explicitly partisan groups. Second, and more unlikely, some interviewees may have chosen to add the research team to a non-partisan group to avoid adding it to the actual partisan thread they maintained.

[9]The data from the app was obtained using the API provided by the app: https://play.google.com/store/apps/details?id=com.narendramodiapp

team of the Prime Minister to spread information about his speeches and party events. Over the past few years, the ruling party (the BJP) has pushed every party member to install the app.[10] The app has been extensively used for political campaigning and to spread the word of the party during elections. We collected data from July 2019 to March 2020 and obtained all posts including text, images and videos that were shared on the platform, focusing on the official posts originating from the team managing the *NaMo* app (rather than from users). This allows us to compare the posts we find on the "local" threads of the ruling party (which constitute the bulk of our data) to the posts made by party executives and other BJP high-level decision-makers.

Coding Instrument and Procedures

Once the data exports were completed, research assistants coded a large portion of the visual content of these threads (roughly 1/9 of the total content; that is, close to 40k images). To determine the subset of data to be coded, we randomly selected a single month of data (out of the nine months of data available) to be coded on *each* of the 533 threads. In each case, we then focused on the visual content posted on the threads.

Two different coders first annotated each sampled image in an interface that reproduced the flow of the threads - while anonymizing the phone number of the participants and other key identifiers. Importantly, this strategy allowed coders to evaluate images in context: while only images were coded, coders were asked to review the text messages that were surrounding the images (either as part of the same post or as adjacent posts) to evaluate them. Figure 2 shows the interface for coding.

Overall, 12 research associates worked on the annotation. A survey of these human coders (more details in Appendix H) shows that they were highly educated WhatsApp users based in Uttar Pradesh, fluent in Hindi and had a clear knowledge of local political dynamics in the state, as evidenced by their responses. Importantly, while they were disproportionately upper-caste, and were also likely to oppose the BJP (the main actor accused of spreading misinformation and hateful content in this context) than to support it. Insofar as opposition party

[10]https://economictimes.indiatimes.com/news/politics-and-nation/amit-shah-wants-1-lakh-download-of-namo-app-in-each-district/articleshow/51735861.cms

supporters frequently accuse the BJP of engaging in the behaviors we attempt to count (misinformation, hateful content, and, partisanship) we see this as the most important dimension on which to have balance.

To ensure that coders followed our coding instructions, they followed a two-days training during which one of the author and a project manager explained coding instructions (enumerated below) in detail, supervised the coding of a mock batch of images and subsequently organized roundtable discussions (between coders) about these mock codings. The objective of this session was less to ensure that any two individuals land on the same coding (which would be impossible, given that codings are indeed *partly* subjective) than to ensure that their rationale was in keeping with our instructions.

If the two coders disagreed on the coding of one of our main concepts (misinformation or hateful content), the image was then forwarded to a third coder, who was tasked with adjudicating. In what follows, our count of misinformed or hateful visuals accordingly refer to materials that were classified as such *by two separate coders*, either in the first instance, or after the intervention of the third coder in case of disagreement between the first two coders.

Given the large amount of material to be coded, we did *not* systematically adjudicate disagreements between coders on our other concepts (for instance, whether the content is arguably "partisan content"). The counts we report on these subsequent concepts accordingly refer to agreements between the first two coders only.[11] [12]

Coding misinformation and hateful content

There is no simple, obvious strategy for human coders to objectively determine whether content should be labeled as "misinformation" or as "hateful content". Accordingly, while we provided

[11]Note, however, that 25% of the disagreements on these other concepts were adjudicated by a third coder. As we show in Appendix F, this did not yield substantially different results.

[12]Though the rest of the paper only focuses on data created by the annotation, we also performed automated analysis on the entire dataset using image clustering techniques to identify and quantify the type of information. Appendix B provides results from this automated analysis.

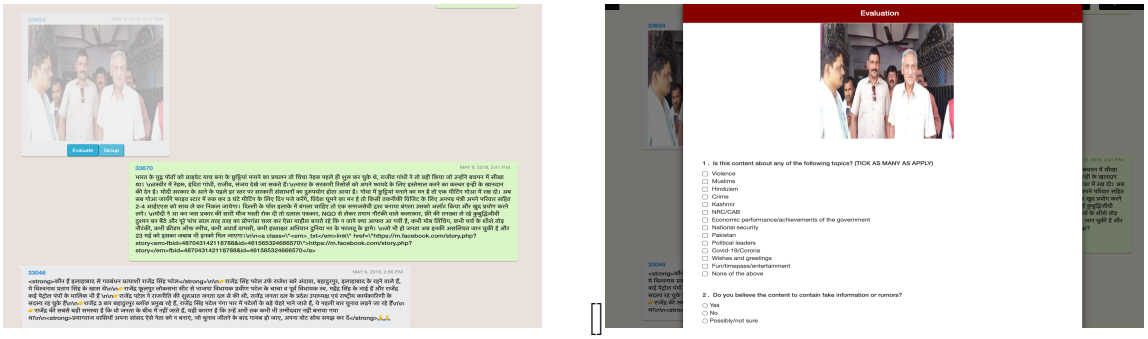

Figure 2. Screenshots of the interface. (a) The entire thread is shown in context, where the annotators pick and image to annotate. (b) Clicking on the 'Evaluate' button on the image opens the annotation workflow.

coders with clear definitions and guidelines as to how to classify content, we eventually rely on inter-coder agreement to evaluate the amount of content labeled as either.

Practically, coders were asked to answer the following four questions about each image:

1. Do you believe the content to contain fake information or rumors? (Yes / No / Possibly - not sure).
2. *(IF response to question 1 above is Yes)*: what is the topic of the fake information? (Muslims / Pakistan / Kashmir / Health/medical advice / Rumor about political leader / Other:).
3. Would you rate the content as explicitly or implicitly hateful or prejudiced towards certain groups? (Yes / No / Possibly - not sure).
4. *(IF response to question 3 above is Yes)*: what is the group targeted? (Muslims / Pakistan / Kashmir / Health/medical advice / Rumor about political leader / Other:).

As they answered the first question, coders were encouraged to check for the veracity of information by consulting the archives of renowned Indian fact-checking websites such as *Altnews* and *BOOM*. As they answered the third question (related to hateful content), coders

were instructed to code as hateful ("Yes") any content that "*promotes discrimination or disparages or humiliates an individual or group of people on the basis of the individual's or group's race, caste, ethnicity, or ethnic origin, nationality, religion, belief system, partisanship, disability, age, sexual orientation, gender identity, or other characteristic associated with systematic discrimination or marginalization.*" In both cases, they were asked to err on the side of caution before answering "No". If they were unable to properly justify such a "No" coding, they were asked to choose the "possibly" response category.[13] This strategy implies that our estimates - presented below - of the amount of hateful content and misinformation present on the threads are *by design* likely *overestimates*.

Given the subjective and interpretative nature of the task assigned to coders (as noted below, professional fact-checkers themselves often disagree), disagreements between the first and second coders were to be expected on a portion of the data.

The chance-corrected inter-coder reliability (Krippendorf's alpha (Krippendorff, 2011)), for the misinformation task was 0.322, for the hate task was 0.323, and for the partisanship task was 0.63.[14] The alpha values for misinformation and hate constitute a fair agreement, and for the partisanship task indicate a substantial agreement (Hughes, 2021). As (Krippendorff, 2011; Artstein and Poesio, 2008) suggest, the exact range is context dependent. Fleiss kappa scores (which are on a similar scale to Krippendorf's alpha Zapf et al. (2016)) measuring inter-annotator agreement between coders for various tasks like hate and misinformation/bias have been in the range of 0.2–0.4 (Lim et al., 2020; Del Vigna et al., 2017; Ousidhoum et al., 2019).

Coding Partisan vs. Nonpartisan Content

In order to better understand how, if at all, partisan threads do advance parties' interests, our second important objective was to evaluate the amount of partisan content in the data, and to classify subtypes of partisan content.

To define partisan content, we asked coders to rely on the following four rules:

[13]This allows us to include "possibly" codings in our analyses below.

[14]Because we had multiple coders who code different sets of images (though every image is coded by at least 2 users), we use Krippendorf's alpha, which is robust to missing values.

- **RULE 1**: an image is deemed "partisan" if a party member OR a party message OR a party logo is either: (a) visible in the image OR (b) not visible in the image, BUT mentioned in the text accompanying the image in a way that makes the image partisan.
- **RULE 2**: a party "message" need not be explicitly attributed to the party. Any idea that is congruent with the party's worldview (for instance, news suggesting that the Muslim population is to take over the Hindu population is here to be considered a partisan message for the BJP). This may be true even if the party is not mentioned or if the party symbol is not shown. Any famous slogan of the party or the opposition (ex: "acche din") also immediately suggests partisan content.
- **RULE 3**: we err on the side of classifying images as partisan. Hence any implication or partisan sight - even if it is subtle - should be classified as partisan. Any image containing an idea or message congruent with partisan ideas should also be seen as partisan.
- **RULE 4**: an image can be partisan either if it is "favorable" to a party/leader/message or if it is unfavorable to another party/leader/message (for instance, an opposition party/leader/message being criticized or mocked).

These rules ensure that we by design count as partisan content a broad range of contents and that we include in that tally any content that might be even just faintly related to partisan matters. As noted below, a number of subsequent questions allow us to further build on these basic answers and provide additional evidence as to the sub-types of partisan content we find in the data.

Quantifying Hateful Content and Misinformation

Using the evaluations of our human coders, we first quantify hateful content and misinformation in the large sub-sample of our WhatsApp data described above.

Coders' Evaluation of Hateful Content

To estimate the amount of hateful content on the threads, we simply calculate the percentage of content that *two* coders rated as "explicitly or implicitly hateful or prejudiced towards certain groups." Since "possibly - not sure" was a possible response to the question, we also provide

estimates that include images coded as "possibly" hateful in the count. Finally, since we can reasonably hypothesize that hateful content might target Muslims in the context we study (the state of Uttar Pradesh in 2019-2020), we also subset our estimates on the group of images that coders separately "tagged" as featuring or having to do with "Muslims."

Table 2: Fraction of images coded as containing hateful content.
(*full data*)

	Hateful	Hateful (*About Muslims*)
Fraction of posts classified as **possibly or clearly** ...	1.9%	21.6%
Fraction of posts classified as **clearly** ...	1.8%	21.5%

Table 2 provides this information for the whole sample. As seen in the first column, only a small fraction (1.8%) of the visual content of the threads was evaluated by *two* coders as *clearly* containing hateful content. The percentage is not much larger (1.9%) if we also include the content that was marked as ambiguous (i.e., as "possibly" containing hateful content) by at least one of the coders. Insofar as coders' evaluations remain — despite the fact that they are based on rules and often, external research — partly subjective, it may lead them to miscode or ignore *some* hateful content. In order to be conservative, it is thus also important to consider the rate of content coded as hateful by at least one of the two coders. This conservative strategy increases our estimates to 5.7%. Subsequent qualitative analysis by the authors of the content classified as "hateful" *by only one of the coders* leads us to believe that this 5.7% number may be a clear upper-bound, since much of this content appears to contain political attacks against rival political leaders (which may disparage them without properly spreading hatred about them).[15] Whether a few percents is a high or dangerous level is however more generally hard to tell: a small fraction of hateful content *might* have dramatically harmful consequences, as

[15]Note, more generally, that our definition of hateful content did cast a potentially wide net, insofar as it included hatred towards individuals in addition to hatred towards entire social groups or minorities.

we suggest below. The mechanism leading to these harmful consequences however matters. This evidence already suggests that any such consequence likely does not owe to a potential "carpet-bombing strategy" which would see users being *mostly* plied with hateful content, as the vast majority of content we code does not fit this category.

These overall estimates however hide meaningful heterogeneity. As mentioned above, one may suspect that hateful content would by definition frequently target the Muslim community. As seen in the second column of Table 2, this is indeed the case. Visuals featuring to or referring to Muslim identity in one way or another[16] were at least 10 times more likely to being coded as containing hateful content by our coders.

Table 3 further breaks down these estimates and specifically focuses on threads associated to the BJP — the bulk of our sample. As seem from the table, the overall amount of content coded as hateful remains as small on BJP threads as on other threads in our sample. However, that hateful content appears disproportionately *more* likely (compared to the estimates presented in Table 2) to target the Muslim community, as seen in the second column. This confirms that an impressive share of the content referring or mentioning muslims on BJP threads *is* hateful content. Besides, as evidenced from the third column, a disproportionate share of that hateful content likely originates from individuals that serve as "admins" on the threads. Given our sampling strategy (focused by design on threads maintained by a party associates), this implies that close party associates and/or party members were much more likely than simple users to disseminate hateful content on these threads. Overall, this confirms that despite its relative scarcity, hateful content disproportionately targets Muslims and disproportionately originates from BJP party associates.

[16]These were identified thanks to a simple "tagging" strategy we describe below.

Table 3: Fraction of images coded as containing hateful content (*BJP subsample*)

	Hateful	Hateful (*About Muslims*)	Hateful (*From Admins*)
Fraction of posts classified as **possibly or clearly** ...	1.6%	37.1%	4.2%
Fraction of posts classified as **clearly** ...	1.6%	37.1%	4.2%

While content is frequently hateful when it focuses on Muslims, it does not follow that most hateful content is about Muslims. Appendix C allows us to understand whether this is the case. As seen from the Figure, while most frequently targeted, Muslims are not the sole target: a large fraction of the content classified as hateful here belongs to a residual ("other") category. As seen in Figure 15 — a word cloud based on response typed in by coders, much of the hateful content indeed appears to target specific partisan individuals or parties. An additional qualitative analysis of a randomly selected sub sample of images (N=50) classified as hateful content targeting "others" confirms this picture: a very large portion of this content (34 out of the 50 images we randomly selected) consists of scathing or frankly threatening messages targeted at political leaders or at specific subgroups within opposition parties, as may be expected on partisan threads. More rarely, this content appears to target a variety of individuals or groups on a non-partisan basis.[17] This implies that hatred along religious or caste lines arguably constitutes a *minor* portion of the overall content we do classify as "hateful content".

Coders' Evaluation of Misinformation

To estimate the amount of misinformation on the threads, we similarly calculate the percentage of content that two coders rated as "containing fake information or rumor." Since "possibly/not sure" was a response to the question (and one we encourage them to use if they could not

[17]Some images for instance target "media", "policemen", "Brahman priests", "pakistan" or "Rohingyas".

justify why a visual was *not* misinformation), we also provide estimates including images coded as "possibly" hateful in the count.

As seen from Table 4, patterns are similar to those described in the previous subsection. Namely, only a very small fraction (2.2%) of the visual content of the threads was evaluated by *two* coders as *clearly* containing misinformation. The percentage is only slightly larger (3.3%) if we also include the content that was marked as ambiguous (i.e., as "possibly" containing misinformation) by at least one of the coders.

Table 4: Fraction of Images Coded as Containing Misinformation.
(Full data)

	Misinformation	Misinformation (*If focused on Muslims*)
Fraction of posts classified as **possibly or clearly** ...	3.3%	8.1%
Fraction of posts classified as **clearly** ...	2.2%	4.1%

Even though coders were encouraged to check content and err on the side of caution before marking it as free of misinformation, they may have failed to acknowledge some misinformation. As above, we thus also consider rate of content coded as misinformation by at least one of the two initial coders. This conservative strategy dramatically increases our estimates (to 5.1%), while overall keeping it at relatively low levels. Qualitative analysis by the authors of the content for which there was disagreement between coders (roughly 600 images) suggests that this higher number may however be a substantial overestimate. This is because a large proportion of this content appears to be akin to jokes (content which do indeed contain false information but which do not *pretend* to represent the truth) rather than deliberate attempts at deceiving users.

As with codings of hateful content, these estimates nonetheless hide some meaningful heterogeneity. Misinformation frequently targets the Muslim community. As seen in the second column of Table 4, this is indeed the case. Visuals featuring to or referring to Muslim identity in

one way or another were significantly more likely to being coded as containing misinformation by our coders.

Table 5 further breaks down these estimates and specifically focuses on threads associated to the BJP. As seen from the table, the overall amount of content coded as misinformation remains as small on BJP threads as on other threads in our sample. However, that content appears disproportionately *more* likely (compared to the estimates presented in Table 4) to target the Muslim community, as seen in the second column. This confirms that a non-negligible share of the content referring or mentioning muslims on BJP threads contains misinformation. Besides, as evidenced from the third column, a disproportionate share of that hateful content likely originates from individuals that serve as "admins" on the threads. This again suggests that party associates disproportionately play a role in the dissemination of this misinformation.

Table 5: Fraction of images coded as containing misinformation.
(*BJP Groups subsample*)

	Misinformation	Misinformation (*About Muslims*)	Misinformation (*From Admins*)
Fraction of posts classified as **possibly or clearly** ...	2.6%	11.6%	5.7%
Fraction of posts classified as **clearly** ...	1.8%	5.5%	5.1%

Misinformation about political leaders constitutes the majority of these images coded as misinformation (about 2/3 of them, in fact, shown in Figure 16 in Appendix D). By comparison, very few actually do focus on the Muslim community (or for that matter, on any of the other topics we had originally listed as potential responses to this question). This is coherent with previously enumerated findings: far more of the activity on these threads targets political leaders from the opposition rather than the Muslim community, even if content targeting Muslims disproportionately tends to contain misinformation or hateful content.

Quantifying Partisan Content

As suggested by the findings detailed in the previous section, the threads in our sample do *not* appear to contain very high shares of misinformation or hateful content. In this section, we in turn try to ascertain whether these threads more generally contain content that may serve the electoral interests of the party they are meant to serve — such as propaganda and/or other information documenting party activities.

Quantifying the Amount of Partisan Content

To do so, we analyze coders' responses to the simple question requiring them to classify each sampled image as "*Clearly not partisan*", "*Not explicitly partisan but referring to national, religious or caste identity (in a way that might be compatible with partisanship)*", "*Possibly partisan*" or simply "*Partisan*". As noted in Section 3, coders were additionally given detailed instructions as to what should count as "partisan" or to one of these other categories.[18] As can be seen from Table 6, coders did not classify a *majority* of the content as partisan or possibly partisan.[19] An even far smaller share (less than 2% of the data analyzed) was additionally earmarked as non-partisan but referring to identity categories that matter politically.[20]

[18]As a reminder, the count we present in this section – and going forward – only rely on agreements between the first two coders. Note though that we validate this strategy in Appendix F, as we show that the intervention of a third coder does not substantially change our results.

[19]Justifications given by coders to classify content as partisan are provided in Appendix Figure 22. As seen from the figure, the most common rationale for coding images as "partisan" owed to the presence of a political actor on the image itself; interestingly, these were however most frequently local political actors such as local leaders or even simple party workers. In keeping with our analyses of the nature of partisan images below, this indicates that much content on partisan threads documents local party activities (meetings, rallies, processions) rather than engage in ideological warfare.

[20]As seen in the table, a very small share of the data is additionally coded as "possibly partisan - not sure". Reasons invoked by coders for classifying images in this category are as such: *Don't understand who/what they are talking about - lack local knowledge to understand*: **13%**; *The topic they discuss possibly has partisan undertones, but I can't tell for sure*: **38%**; and *Other*: **49%**.

Table 6: Fraction of images coded as containing partisan content.

	Full Sample	BJP subset
Fraction of posts classified as partisan **OR** possibly partisan.	41.3%	41.2%
Fraction of posts classified as partisan, **OR** possibly partisan, **OR** referring to identity groups.	42.9%	42.7%

The prevalence of non-partisan content holds whether or not we restrict the sample to BJP threads, which may be hypothesized to be better managed or organized. In line with our coding criteria, this implies that *most* of the visual content did not feature a party member, a party message or a party logo. And additionally, given our definition of what might count as a "party message", that it did not convey or imply a message that could conceivably be seen as either congruent to the party's ideas or even just critical of opposition parties. This is remarkable, insofar as it comes to show that only a minority of the content posted on threads administered by party associates, presumably for political and electoral reasons, is in fact political in nature. This may indicate that these threads are not meant to disseminate political or partisan content, or alternatively, that their admins do not manage to maintain the focus of the threads on politics, because users post other types of content.

Figure 3 shows that the truth lies somewhere between these two explanations. As can be seen from the Figure, BJP admins post considerably more partisan content than users, who post partisan content only around a third of the time. This hints to the fact that admins do not control the type of content that users choose to post, and hence likely struggle to keep the threads as focused on politics as they might want to. Since admins themselves post politically relevant content only 65% of the time, this however also suggests that they themselves diversify content. The fact that admins on groups explicitly dedicated to the party post roughly 30% of non-partisan content implies that these group threads are social spaces in addition to (or maybe even *more than*) simple diffusion lists for party propaganda. This may in turn be because admins believe their partisan messages will be more impactful if they include it in a flow of other content (for example salutations and wishes), or because they less consciously post such

content. Importantly, the posting behavior of elite posters on the NaMo App largely mimics the diversification of content of users (only 1/3 of NaMo content is classified as partisan by our coders), which lends credit to the idea that party elites feel the need to see diverse content on partisan social media.

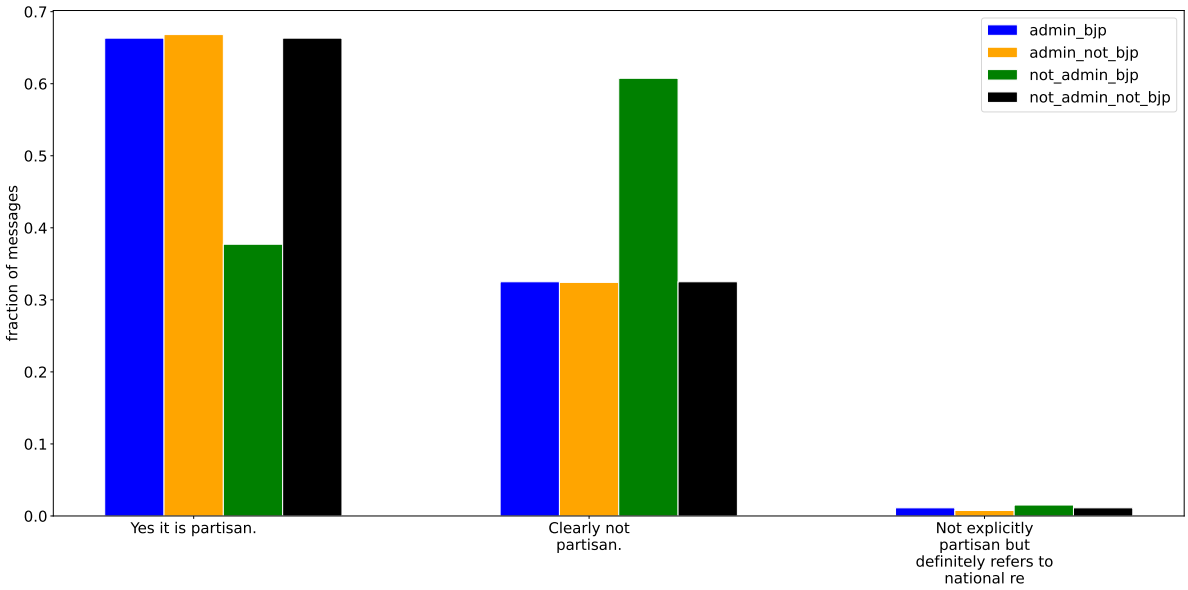

Figure 3. How partisan is the content of the images?
(by whether the poster is an admin or not, and BJP or not.)

What Type of Partisan Content Appears on Partisan Threads?

Partisan threads thus contain remarkably fewer partisan visuals than their denomination might suggest. Does that content however serve the electoral interests of parties, help propagate their ideas, or contribute to the *persuasion* of voters, as recurrent media narratives suggest they do?

One way to touch on this question is to evaluate the amount of content classified as partisan (a broad concept, as per our definition) that resembles political advertising or other forms of strategic political communication implicitly or explicitly carrying an idea or message (for instance, political memes). Even if political scientists have repeatedly shown political advertising to have small or nonexistent effects on political preferences, regardless of the type of

content they include — whether it is negative or promotional[21], and regardless of the characteristics of voters themselves (Coppock et al., 2020), the presence of such content on threads would at the very least suggest that admins, and potentially other posters, *aim* to disseminate some type of party message. This may not apply to other, less strategic forms of content which may also fit our definition of "partisan". Frequent examples in our data include photos of local politicians on the occasion of their birthday or selfies by party supporters attending a political event - categories we return to below. While they may serve different purposes (organizational, for instance), such contents cannot easily be assumed to help spread a party *message*, let alone *persuade* voters.

Accordingly, we use in this section additional codings performed on the subset of sampled content classified by our human coders as "partisan" to better describe its nature, and hence to reflect on the function that these partisan threads might serve. A first, admittedly coarse way to identify whether the visual content present on threads contains a message of some kind is simply to evaluate the extent to which this content contains *text*, either directly on the image (as in the case of partisan memes) or in a adjacent text message directly related to and accompanying the image.[22] As seen from Figure 23, a very small share (far less than 10%, altogether) of the content we coded as partisan did contain *any* kind of text at all, whether on the image or accompanying it. This already disproves the idea that partisan threads enable the mass dissemination of partisan propaganda. Altogether, this suggests that only a few percents of the visual content contained on these threads contains a partisan message *in writing*. This in turn implies that the overwhelming majority of the content we classify as "partisan" is less effective or more ambiguous than it could ideally be at spreading the party line. We accordingly take this as a first sign of the relatively low quality of the partisan content posted on partisan threads.

[21]As seen in Figure 27, the content we code as partisan rarely tends to be overwhelmingly positive content. Neutral content here refers to content whose negative or promotional nature coders cannot determine. A qualitative analysis of that content shows that this sub-sample of our data includes photos of political events or political workers, most often posted there without any comment.

[22]As a reminder, our coders do see images in their original context — that is, they see the whole thread, including messages above and below images, as they are coding these images.

What is more, as seen in Figure 23, quotes from party leaders — probably the most direct way to think of how party messages may be disseminated — only constitute a small fraction of the already very limited quantity of partisan text that we find on or around visuals on partisan threads. Unattributed text and indirect statements may in practice achieve the same effect and are slightly more frequently found on threads. Taking into account the fact that less than 10% of all partisan content contains text, this however remains a very small proportion of all content. In addition, a qualitative analysis of the data classified in Figure 23 as "other" suggest that the text if often merely descriptive of the actions of leaders, as in "politician X did this yesterday" or "politician Y visited that today", rather than in making some argument about the qualities or weaknesses of the individuals portrayed. What is more, the leaders described in these visuals are far more frequently local (district or block-level) party leaders than national leaders. This disproves the idea that these threads may be contributing to the personification or the "presidentialisation" of Indian politics: at the local level, most of the attention remains pointed at local politicians.

Figures 24 and 25 (Appendix F) further confirm that the partisan content posted on these threads may overall be low quality, from the standpoint of a political party trying to convince or persuade voters. In Figure 24, we specifically try to evaluate the extent to which the partisan content that contains text is akin to the memes on which much of the BJP strong social media game has been built (Udupa, 2018, 2019). As described in Sinha et al. (2019), memes mocking opposition leaders and/or praising Narendra Modi have constituted an essential part of the digital efforts of the top brass of the party. In Figure 25, we however show that these official, attributed memes constitute a *minor* part of the content posted on threads. Worse maybe from the party leadership perspective: admins and party elites posting on NaMo do not perform much better than regular users on this metric. This once more points to the fact that the type of content pushed by party higher-ups – whether or not that content is actually "effective", which we cannot evaluate from these data alone – tends to get diluted in other contents, including by the admins of threads themselves.[23] Figure 26 in turn shows that much less of the content posted on threads (compared to NaMo) is unsourced and unattributed,

[23]This is coherent with the main descriptive finding of (AUTHOR 2021). Namely, that the ground-level "social media workers" of parties, including the BJP, are on average very undisciplined and inefficient disseminators of party propaganda.

which suggests that threads contain very different content than what party higher-ups would ideally like to see pushed on to users.

A second strategy to qualitatively assess partisan visuals posted on threads is to understand what the individuals depicted in these visuals (as noted above, predominantly *local* politicians) concretely do in these images, whether or not these visuals do also contain text. This is what we do in Figure 26. Images depicting different activities may have different functions or be useful to parties in different ways. Establishing what each type of image might achieve is overall unclear, and in any case beyond the purview of this article; it can however safely be assumed that *some* types of images contribute more than others to the dissemination of the party message, or to the development of the image of the party or of its candidates. Images depicting a politician engaging in a good deed (charity and/or public distributions) or interacting with constituents around a public function may help convince voters of the worth or the dedication of that individual. Similarly, images showing that a local politician is meeting party higher-ups may provide evidence of their "upward connectedness" (Auerbach and Thachil, 2018) and hence of their ability to solve problems; Visuals of political actors participating to a political event (a rally, a meeting etc.) may help show that a politician is active and/or "viable" electorally, but it is similarly unlikely to do much to develop that individual's image. Simple portraits of politicians doing nothing specific may achieve even less from this point of view.

Figure 26 informs this discussion and in our view confirm our central finding — that the "partisan" content (already a minority of the content posted on these threads) often does very little to build the image of candidates or their parties in an electorally meaningful manner. As seen from the figure, the subset of visual partisan content featuring a political actor frequently shows that actor as doing "nothing specific" (as in the case of simple portraits). A larger part of the content shows these actors participating to various political events and/or meeting other political actors, which may be more useful electorally, albeit in a more indirect manner than straight propaganda might. Most importantly, only a very small share of this content actively displays a political actor engaging in a good deed of the type that may be more easily leveraged in the context of electoral campaigns. Combined with the insight that the political actors displayed in these visuals often are local politicians or simple political workers (selfies of supporters/workers attending meetings are in our dataset a large category), this confirms that

the content of these threads infrequently push the party line (as opposed to local politicians' interests), and more generally, that it may not be optimally useful for persuasion or propaganda purposes.

The comparison between the content posted on our local threads and the content posted on NaMo further increases our confidence that these threads' content is underwhelming from a party propaganda point of view. As seen in the figure (26), the content featured on NaMo and on local threads is *extremely* different. While much of the content on local threads shows actors participating to political events and/or meeting other political actors, this type of content is almost entirely absent on NaMo, which likely reflects the fact that NaMo focuses on national-level politics while local threads consistently do not. NaMo also emerges as containing more content akin to straight-out propaganda (see "good deed" category). Finally, the NaMo content far more often fits into our "nothing" and "other" categories.[24]

Beyond Partisan Content, What Appears on Partisan Threads?

We rely on additional codings to answer this question and describe the 50% or so of the visual content of threads that do not fit our broad definition of "partisan content".

Classifying Non-partisan Content

We first break down this content according to categories that we ex-ante expected to be common, based on a limited qualitative analysis of 20 threads. As shown in Figure 4, a number of these pre-determined categories emerge as particularly frequent[25]: wishes, salutations, greet-

[24]Qualitative analyses of this content readily explain why this would be the case: NaMo images comparatively contain a far larger number of memes as well as other visuals containing a clear written content. Political leaders featured in these images (mostly Narendra Modi, but also opposition leaders in the case of more negative memes) are indeed often doing "nothing special" in these images, as the image merely is an official portrait of them. In other cases, they are actually performing an action, for instance performing a phone call or praying - hence the "other" coding.

[25]The figure only provides estimates for categories that returned $> 1\%$ of the time. In addition to the categories labeled on the figure, coders could also choose among the following: Song (Non-religious)/Tiktok videos/Food pictures/ Sexual content - porn - nudity (without

ings and to a lesser extent prayers (and other religious messages) together account for over one third of the content posted by the admins of BJP threads, pointing to the importance of phatic content (Berriche and Altay, 2020) in group-based discussion apps.

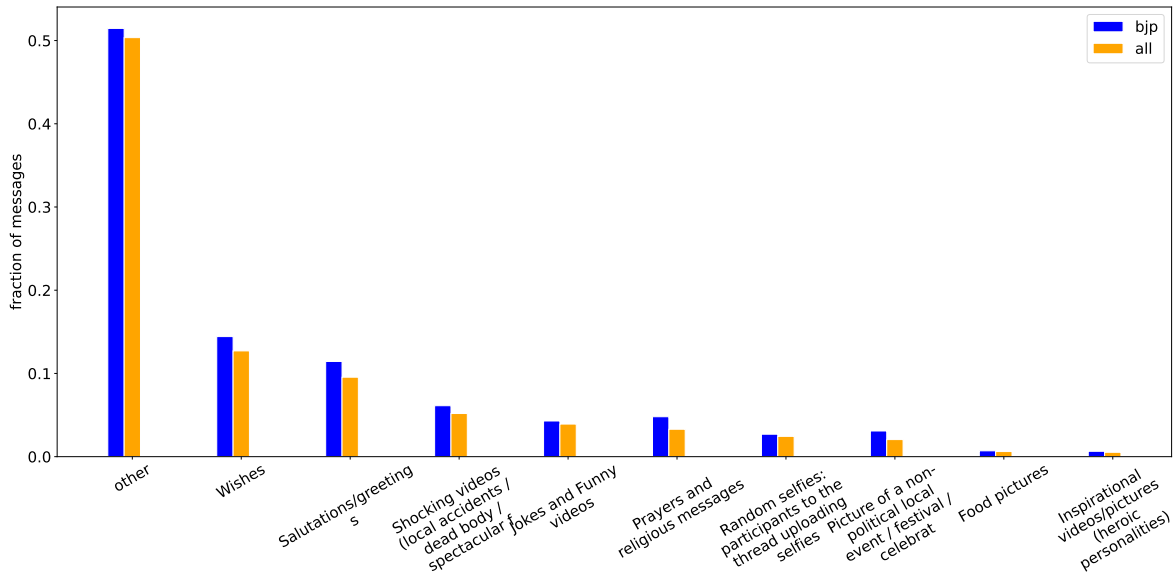

Figure 4. If it is *not* partisan content, what kind of content is it?

Besides, a follow-up analysis carried in the Fall of 2021 of the content classified under our residual category ("other" in Figure 4) confirms that the non-partisan content largely consists of content that does not easily lend itself to political or partisan propaganda.

As shown in Figure 5, a *majority* of this visual content consists of text-free photos whose appearance suggests they were captured from a phone camera (potentially the camera of the poster himself, though it may also be forwarded or borrowed content), as opposed to memes, screenshots of articles, or images of print newspaper, all of which more easily lend themselves to a type of messaging. Besides, as shown in Figure 6, a very large part of this content appears to have been used to document local gatherings, events or remarkable sights, or even to expend the reach of official documents and notices in paper forms. In that sense, local partisan WhatsApp threads appear to double down as local notice board, as users flood

any political content or implication) / Stunts / Picture of a non-political local event/Other Non-political news.

their content with local information of public interest. More problematically maybe, in light of fears around the role these threads may play in vigilantism, a non-trivial amount of this content relates in one way or another to the denunciation of suspected crimes or criminals.

Additional questions about the content that coders originally classified as "other" suggest that they may have failed to *not* classify the many "good morning" messages that appear on threads under "salutations/greetings."[26], and hence that threads may contain even more phatic content than what Figure 4 already suggested. As shown in Figure 5, around 15% of this content consists of memes, the vast majority of which include either wishes, good morning messages or motivational sayings, such as the one presented in Figure 7 (Figure 32 presents a full breakdown of the types of memes we detected in this subset of the data). By contrast, few of these memes appear related to issues/topics that may lend themselves to an implicit type of political messaging.

[26]They treated these as specifically addressed to someone, as opposed to the more general, less attributed "good morning" messages we describe here.

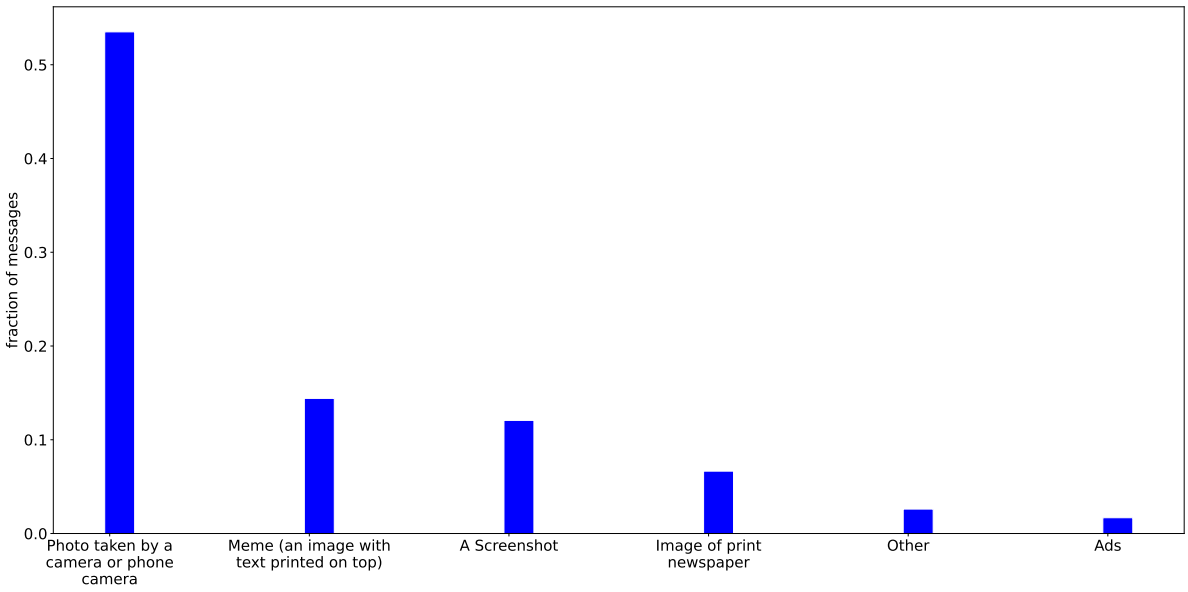

Figure 5. What type of content did we classify as "other" in Figure 4?

Social Identities on Threads

In addition to this general typology, we also ask our coders to specifically evaluate whether the content - both partisan and non-partisan - refers to national, religious, caste or any other social identity. Further, we ask them, using pre-determined categories, to describe the subset of content they do classify as containing a reference to social identity categories.

Overall, coders marked **21%** of the content of the threads as containing visual and/or textual references to social identities. In line with our findings above, much of this content focuses on Hindu content. As seen from Figure 8, almost 3/4 of this content arguably is "Hindu content". Visuals featuring pictures of gods and goddesses are especially prominent in this. Altogether, this suggests that almost 15% of all images posted on partisan threads allude to Hinduism, more or less directly.

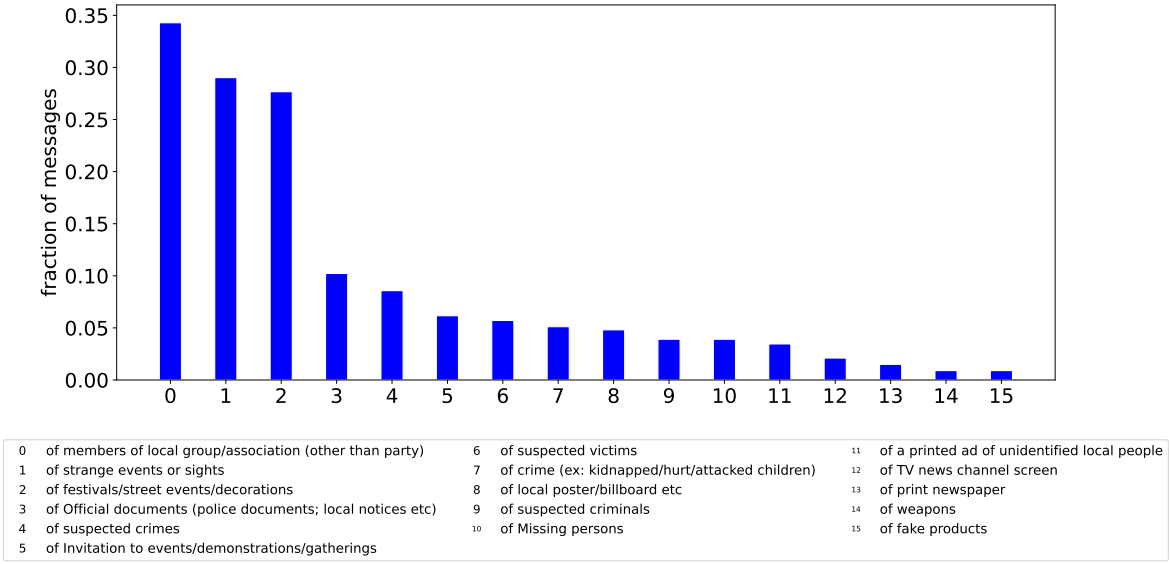

Figure 6. What are Photos taken by phone camera of?

Figure 7. Example of a "Good Morning" Image.

While little of this content actually falls under partisan content[27], religious posts may serve an indirect partisan function, since the BJP (from which the bulk of our data originates) self-defines as a Hindu party. Accordingly, even if partisan content overall rarely appears to help disseminate a partisan message, it maybe that these threads serve another function and achieve it thanks to other types of content. While exploring the effect of repeated exposure to religious (Hindu) materials is beyond the scope of this article, it is for instance easy to hypothesize that this content would help build up the "Hindu" category, eventually a politically constructed category. In that sense, threads may be useful to the party's top brass in a much more indirect manner.

[27]Only 7% of the posts labeled as one of *'Religious Hindu salutations/greetings'*, *'Pictures/videos of gods/godesses'*, *'Hindu prayers/devotional material/songs'* or *'Hindu pride messages'* are labeled as partisan.

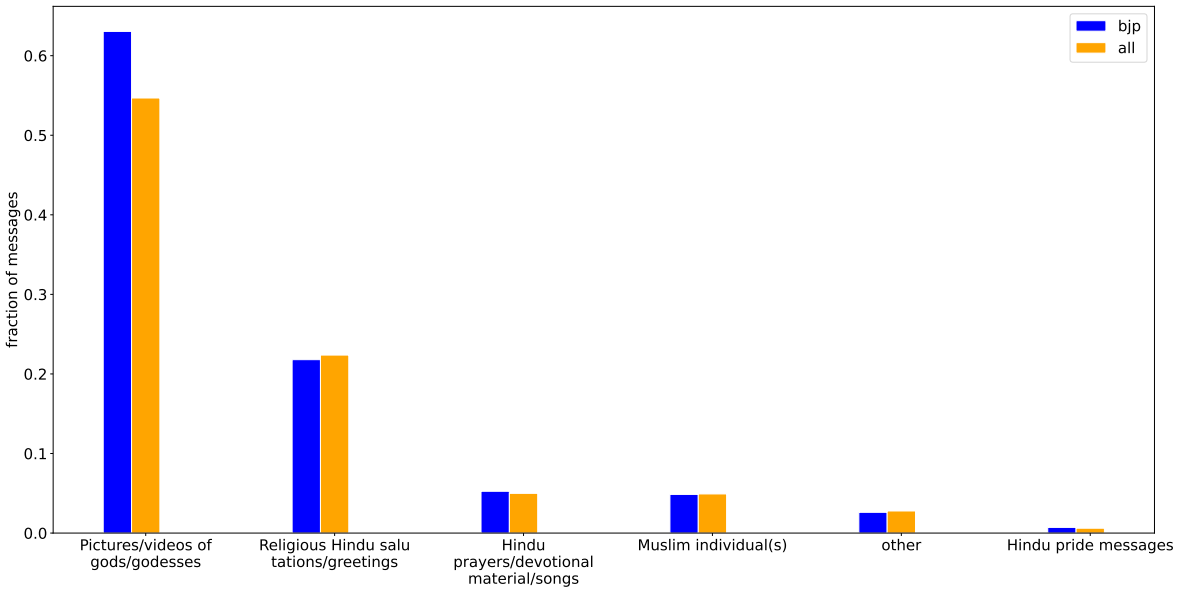

Figure 8. If the content of the image refers to national/religious/caste identity, how does it do so?

Discussion

Overall, our analyses suggest that partisan threads (including BJP threads, the bulk of our sample) contain relatively few hateful or misinformed posts, and more surprisingly maybe, that *most* content cannot easily be classified as "partisan content". Qualitative analyses confirm our quantitative intuitions. Salutations and wishes, often formulated in religious terms and/or relying on religious iconography, constitute much of the content. A large share of the content is also neither partisan nor religious, and more easily classifiable as phatic (Berriche and Altay, 2020) or entertainment-related. Finally, threads contain a vast amount of photos and information whose goal really seems to be to inform users of (non-partisan) local events, gatherings or to inform them in some other way. As we have seen above, this type of "notice board" content is largely driven by users (i.e. participants we know not to be "admins").

Additional data and analyses will be necessary to understand why partisan elites rely so heavily on WhatsApp and what the effect of exposure to these threads might be. While descriptive findings such as these ones do not allow us to answer these questions, these findings

are however not easy to reconcile with the common narrative about partisan uses of WhatsApp in India that motivated us to look into these threads in the first place.

While they contain some misinformation and some hatred (especially towards Muslims), it is probably inexact to say that partisan WhatsApp threads are *packed* with misinformation and hateful messages; similarly, while they do contain partisan content that may seem valuable from an electoral persuasion perspective, they also do contain many other types of partisan content (for instance, party workers selfies and birthday wishes to political actors). All of these different types of partisan content float in a bath of unrelated, often entirely non-political content. Much of this non-political content owes to simple users, who appear to use the threads as notice boards, and/or to post phatic content. Even if one were to argue (reasonably, in our opinion) that content framed in religious terms and/or leading to increased religious polarization also serves an indirect partisan function, it probably remains wrong to think of WhatsApp groups as a propaganda pipeline. Unsurprisingly maybe, given the relatively horizontal architecture of WhatsApp (admins are not very powerful and cannot easily prevent content from being posted), partisan WhatsApp groups are first and foremost partisan *in name*. They are communities of individuals assembled by a party agent. While they likely contain more partisan content than threads formed with a different pretense or by a non-partisan actor, they may not be the electoral silver bullet that they have sometimes been made out to be, including by political leaders themselves (who may have a reason to want us to believe this).

These findings should lead us to address new questions in future works. Two interrelated questions in our view deserve further exploration. The first one concerns the motivations of political actors. In light of what we now know of the content of these threads, the motivations of both party leaders and local-level actors should be puzzling. Why are party leaders so keen to organize pyramidal WhatsApp structures, if local-level threads eventually only provide a messy tool for the dissemination of party messages? Is it because they systematically overestimate the ability of these structures to serve party interests, or alternatively, because they believe these threads to fulfil a different function? One possibility, consistent with parts of our evidence, is that partisan WhatsApp groups mainly have an organizational function. Rather than contribute to the dissemination of party propaganda targeting voters, it may be that these threads mostly target party associates and sympathizers, providing them with a useful platform to communicate on party events and/or exchange among themselves, as in any other

organization. The relative abundance of party workers selfies on these threads – a phenomenon visually documented in Appendix B – would be consistent with this explanation.

Second, to what extent, and how, might information and misinformation circulating on these threads – in the contexts we have described here – contribute to attitudinal and behavioral changes? Our findings may be naively interpreted as suggesting that these effects should be disappointing or minimal, in light of the (low) quality of the partisan content posted on local threads. The relative scarcity of problematic content that we highlight may by the same token suggest that the partisan WhatsApp misinformation problem that has become a PR nightmare for the platform may have been overblown. It is however important to reiterate here that different types of analyses will be needed in order to evaluate the causal effects of *exposure* to these different types of content, and that we as a result remain entirely agnostic as to the effects of exposure to these contents, in the context we have described here. Our results must nonetheless lead us to speculate as to the way in which the WhatsApp architecture – and more broadly, the architecture of closed private discussion apps – could enhance or decrease the believability of information and misinformation.

Practically, this should lead us to wonder whether the messy and horizontal structure of these groups may also be particularly dangerous for a different reason: because misinformation disseminated through this architecture may be more easily believed, regardless of its rarity. A small number of problematic posts, even if they are drawn in a flux of other content, may achieve a spectacular effect; similarly, a small number of savvy partisan posts may help build an electoral victory. Our descriptive results allow us to better understand the context in which these types of content may appear. They clearly show that this *potentially* impactful content is largely drawn in a mass of other, largely-unrelated content. This is not surprising given the relative inability to control content that "admins" have on WhatsApp groups. We however also see this as suggestive of a more complex narrative.

Following this line of argument, closed partisan threads may not be powerful because they unleash a barrage of misinformation or propaganda on users in a coordinated manner, but precisely because these contents are interspersed between seemingly innocuous content. If this "other" content helps users develop a sense of community and/or trust, the potential effect of either partisan posts or misinformation may be multiplied. The widely entertained influence

and power of closed discussion groups may, in other words, lie less in their ability to distribute large quantities of information at a very low cost than in their ability to expose patiently constructed - and trustful - digital communities to selective misinformation and propaganda.

Acknowledgments

This study received clearance from Leiden University's data protection officer in March 2019. The authors thank Ved Prakash Sharma and his team for excellent research assistance.

References

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Auerbach, A. M. and Thachil, T. (2018). How clients select brokers: Competition and choice in india's slums. *American Political Science Review*, 112(4):775–791.

Avelar, D. (2019). Whatsapp fake news during brazil election 'favoured bolsonaro'. *The Guardian*, 30.

Badrinathan, S. (2020). Educative interventions to combat misinformation: Evidence from a field experiment in india. *Working Paper*. https://sumitrabadrinathan.github.io/Assets/FakeNewsPaper.pdf.

Badrinathan, S. and Chauchard, S. (2021). "i don't think that's true, bro!"an experiment on fact-checking misinformation in india. *manuscript*.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., and Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, 26(10):1531–1542.

Berriche, M. and Altay, S. (2020). Internet users engage more with phatic posts than with health misinformation on facebook. *Palgrave Communications*, 6(1):1–9.

Cheeseman, N., Fisher, J., Hassan, I., and Hitchen, J. (2020). Social media disruption: Nigeria's whatsapp politics. *Journal of Democracy*, 31(3):145–159.

Coppock, A., Hill, S. J., and Vavreck, L. (2020). The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments. *Science Advances*, 6(36).

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., and Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. In *In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*.

Evangelista, R. and Bruno, F. (2019). Whatsapp and political instability in brazil: targeted messages and political radicalisation. *Internet Policy Review*, 8(4):1–23.

Farooq, G. (2017). Politics of fake news: how whatsapp became a potent propaganda tool in india. *Media Watch*, 9(1):106–117.

Flynn, D., Nyhan, B., and Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, 38:127–150.

Fourney, A., Racz, M. Z., Ranade, G., Mobius, M., and Horvitz, E. (2017). Geographic and temporal trends in fake news consumption during the 2016 us presidential election. In *CIKM*, volume 17, pages 6–10.

Garimella, K. and Eckles, D. (2020). Images and misinformation in political groups: evidence from whatsapp in india. *Harvard Kennedy School Misinformation Review*.

Garimella, K. and Tyson, G. (2018). Whatsapp doc? a first look at whatsapp public group data. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Gil de Zúñiga, H., Ardèvol-Abreu, A., and Casero-Ripollés, A. (2019). Whatsapp political discussion, conventional participation and activism: exploring direct, indirect and generational effects. *Information, communication & society*, pages 1–18.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. (2019). Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378.

Guess, A., Nagler, J., and Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances*, 5(1):eaau4586.

Guess, A. M., Lockett, D., Lyons, B., Montgomery, J. M., Nyhan, B., and Reifler, J. (2020a). "fake news" may have limited effects beyond increasing beliefs in false claims. *Harvard Kennedy School Misinformation Review*, 1(1).

Guess, A. M., Nyhan, B., and Reifler, J. (2020b). Exposure to untrustworthy websites in the 2016 us election. *Nature human behaviour*, 4(5):472–480.

Hughes, J. (2021). krippendorffsalpha: An r package for measuring agreement using krippendorff's alpha coefficient. *arXiv preprint arXiv:2103.12170*.

Jerit, J. and Zhao, Y. (2020). Political misinformation. *Annual Review of Political Science*, 23:77–94.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Lim, S., Jatowt, A., Färber, M., and Yoshikawa, M. (2020). Annotating and analyzing biased sentences in news articles using crowdsourcing. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1478–1484.

Lokniti, C. (2018). How widespread is whatsapp's usage in india?

Lupu, N., Bustamante, M. V. R., and Zechmeister, E. J. (2020). Social media disruption: Messaging mistrust in latin america. *Journal of Democracy*, 31(3):160–171.

Newman, N., Fletcher, R., Kalogeropoulos, A., and Nielsen, R. K. (2019). Reuters Institute Digital News Report 2019 . Reuters Institute for the Study of Journalism.

Nuraniyah, N. (2019). The evolution of online violent extremism in indonesia and the philippines. *Global Research Network on Terrorism and Technology Paper*, (5):1–17.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., and Yeung, D.-Y. (2019). Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4667–4676.

Pennycook, G., Cannon, T. D., and Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, 147(12):1865–1880.

Perrigo, B. (2019). How volunteers for india's ruling party are using whatsapp to fuel fake news ahead of elections. *TIME*. January 25, 2019. https://time.com/5512032/whatsapp-india-election-2019/.

Purohit, K. (2019). Post caa, bjp-linked whatsapp groups mount a campaign to foment communalism. *The Wire*. December 18, 2019. https://thewire.in/media/cab-bjp-whatsapp-groups-muslims/.

Reis, J. C., Melo, P., Garimella, K., Almeida, J. M., Eckles, D., and Benevenuto, F. (2020). A dataset of fact-checked images shared on whatsapp during the brazilian and indian elections. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 903–908.

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. In *The World Wide Web Conference*, pages 818–828.

Rossini, P., Stromer-Galley, J., Baptista, E. A., and Veiga de Oliveira, V. (2020). Dysfunctional information sharing on whatsapp and facebook: The role of political talk, cross-cutting exposure and social corrections. *New Media & Society*, page 1461444820928059.

Schakel, A. H., Sharma, C. K., and Swenden, W. (2019). India after the 2014 general elections: Bjp dominance and the crisis of the third party system. *Regional & Federal Studies*, 29(3):329–354.

Sinha, P., Sheikh, S., and Sidharth, A. (2019). *India Misinformed: The True Story*. HarperCollins India, Noida.

Treré, E. (2020). The banality of whatsapp: On the everyday politics of backstage activism in mexico and spain. *First Monday*.

Udupa, S. (2018). Enterprise hindutva and social media in urban india. *Contemporary South Asia*, 26(4):453–467.

Udupa, S. (2019). Nationalism in the digital age: Fun as a metapractice of extreme speech. *International Journal of Communication*, pages 3143–3163.

Vaccari, C. and Valeriani, A. (2018). Digital political talk and political participation: Comparing established and third wave democracies. *SAGE Open*, 8(2):2158244018784986.

Valeriani, A. and Vaccari, C. (2018). Political talk on mobile instant messaging services: a comparative analysis of germany, italy, and the uk. *Information, Communication & Society*, 21(11):1715–1731.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.

Wittenberg, C. and Berinsky, A. J. (2020). Misinformation and its correction. *Social Media and Democracy: The State of the Field, Prospects for Reform*, page 163.

Zapf, A., Castell, S., Morawietz, L., and Karch, A. (2016). Measuring inter-rater reliability for nominal data–which coefficients and confidence intervals are appropriate? *BMC medical research methodology*, 16(1):1–10.

Zizumbo-Colunga, D. and del Pilar Fuerte-Celis, M. (2020). The political psychology of lynching: Whatsapp rumors, anti-government appeals, and violence.

APPENDIX A: Justification for "Public Interest" Criterion With Respect to GDPR Rules.

In this appendix, we make a case for why the research we undertake here arguably is in the public interest.

As explained in the body of the article, it is in this configuration impossible, from a practical standpoint, to obtain each participant's consent. The question is thus why researchers should be allowed to go ahead with the analysis of the full content of the threads, including the material posted by participants whose consent has *not* been explicitly obtained. Note that having access to the thread most often means that we have access to their phone number. This leads us to the following question, related to the aforementioned exception to GDPR rules: why is the proposed research project arguably in the public interest?

The main rationale as to why analyzing the content of threads created and maintained by political parties – public actors – is in the public interest in our view relates to the likely content of these threads, and to the need for independent researchers (that is, researchers not affiliated to WhatsApp nor to the Indian State) to document this content in as unbiased as possible a manner.

In a nutshell, these threads are suspected of containing content that violates a number of criminal acts (both in India and in Europe), as documented below. Acknowledging and evaluating the presence of such reprehensible and criminal content is arguably in the public interest as it seems necessary to document and weigh the scope of a criminal activity in order to address it adequately. In parallel, it should be noted that none of the main stakeholders (either the platform, WhatsApp, and the Indian state) are likely to carry this task, insofar as neither of these actors has much to gain from doing so. The rest of this memo expands on each of these points.

India has over the past few years emerged as a hotspot in the global misinformation crisis. In recent years, messages containing misinformation about electoral security, terrorism, HIV/AIDS, and vaccine safety have circulated widely in the country, arguably leading to a slay of problematic behaviors offline (murders, riots, and other harmful behaviors), and possi-

bly affecting the outcome of elections (Sinha et al 2019). As argued by the Indian state itself, both partisan and non-partisan threads often contain fake or erroneous news, many of which are suspected to lead to offline crimes punishable under a variety of articles of the Indian penal code. Mob lynchings incidents triggered by hateful messages circulated on the platform have especially attracted attention, both in India and abroad. These cases have been well documented in prominent Indian and international press outlets[28], with some observers referring to misinformation as a major "public health crisis".[29]

This has led the Indian state to ask that WhatsApp deencrypts the data contained on threads[30], which the platform has so far, for better or worse, refused to do. This points to the fact that the platform itself, WhatsApp, is extremely unlikely to investigate and document the existence of criminal activity on threads, which raises a major public policy question (as well as a question about GDPR rules), as it may not be desirable from a political or legal point of view to create privacy rules that protect criminal activities from being screened and evaluated. Significantly, Indian political parties (especially the ruling party, the BJP) have often been accused to play a central role in the diffusion of this misinformation (Sinha et al 2019), with some analysts accusing the ruling party of deliberately spreading divisive and legally reprehensible misinformation ahead of elections.[31] Political parties have indeed over the past decade developed impressive networks to diffuse information, and according to a number of aforementioned journalistic reports, misinformation.[32] As suggested by each of these references,

[28]https://www.bbc.com/news/world-asia-india-45140158 https://time.com/5512032/whatsapp-india-election-2019/
https://www.ndtv.com/india-news/whatsapp-rumours-updates-over-whatsapp- rumours-5-men-killed-in-maharashtra-men-beaten-in-chennai-1876287

[29]https://www.nytimes.com/2019/04/29/opinion/india-elections- disinformation.html

[30]https://www.indiatoday.in/technology/news/story/whatsapp-maintains-its- stand-on-govt-s-request-for-message-traceability-in-india-1551098-2019-06-18

[31]https://qz.com/india/1534754/modis-namo-app-spreads-pro-bjp-fake-news- before-indian-elections/
https://www.theatlantic.com/international/archive/2019/04/india- misinformation-election-fake-news/586123/

[32]https://www.huffingtonpost.in/2017/09/11/bjp-may-have-created-a-monster- with-its-troll-army-but-amit-shah-understands-it-may-turn-on-them-one- day$_{a2}$3204198/*https* : *//www.hindustantimes.com/india − news/bjp − plans − a − whatsapp − campaign − for −* 2019 *− lok − sabha − election/story − lHQBYbxwXHaChc*7*Akk*6*hcI.html*

the ruling party (the BJP) emerges in a wide array of press reports as the main suspect in this party-led misinformation crisis. Insofar as the BJP is currently in office, this points to the fact that the state itself is unlikely to investigate and document misinformation on the platform in an unbiased manner, as it would have much to lose to doing so. It should in addition be noted here that such involvement by Indian authorities – insofar as it would not be subject to GDPR rules – may be normatively undesirable from a privacy standpoint.

All of this in our opinion points to the need for independent, unaffiliated researchers to investigate this issue. Knowing the scope and the type of (mis)information circulating on party threads would greatly contribute to the growing global debate over the privacy and encryption of discussion apps around the world.[33]

If it was found, as can be suspected, that a massive amount of misinformation, hate speech, or other messages inciting to violence circulate on party threads, whether these messages originate from actors affiliated to the party or whether they are merely left uncorrected by partisan actors, it may suggest that platforms need to step up their efforts to control content. It may also affect the role assigned to states in this discussion. In the Indian case, it is hard to see why de-encryption would be the solution to the misinformation problem if much misinformation originated from the ruling party threads. It may instead suggest that political parties – public actors in a democracy – need to be held accountable for the content they encourage or openly disseminate.

For all these reasons, it appears to us that investigating this issue would be in the public interest – acknowledging of course that an extensive array of protocols was additionally put in place to protect the identity of thread participants.

[33]Significantly, the US government has been debating this issue, as has the EU, and the Indian government:
https://www.forbes.com/sites/zakdoffman/2019/06/29/u-s-may-outlaw-uncrackable-end-to-end-encrypted-messaging-report-claims/
https://carnegieendowment.org/2019/05/30/encryption-debate-in-european-union-pub-79220
https://www.loc.gov/law/help/encrypted-communications/european-union.php
https://www.vox.com/2019/2/19/18224084/india-intermediary-guidelines-laws-free-speech-encryption-whatsapp

APPENDIX B: Descriptive Analysis using Automated Methods

Where possible, automated computer vision techniques were used to complement human codings and allow us to quantify the prevalence of some types of content *in the full dataset.*

We used two types of image clustering techniques, which serve different purposes. First, we used a perceptual hashing technique (Garimella and Eckles, 2020) to identify perceptually similar images. The perceptual hashing algorithm works by randomly sampling patches of the image to obtain a hash value. The property of such a hash is that similar images have a similar hash value. These are images which are near duplicates of each other but have minor modifications like simple cropping, or an additional water mark, or minor changes in text on the image.

After we obtain the hashes (an alpha numeric string), due to the property that similar images have similar hashes, we can define define a distance between them. We used Hamming distance between the hashes in this paper. Next, based on these distances, we used a clustering algorithm called DBScan (Ester et al. 1996) to cluster similar images. An example of such images identified from our data is shown in Figure 9. As can be seen here, this strategy allows us to group near similar images.

Figure 9. Near similar images obtained using a clustering of perceptual hashes. The three images seem to be taken back to back and involve minor differences, e.g. a man in black t-shirt on in the left in for image (a).

Since the aforementioned perceptual hashing algorithm used above only identifies near duplicates of the same image, it does not identify semantically similar images. To identify

semantically similar images, we used a different algorithm. Using a pre-trained convolutional neural network (CNN) trained with Facebook's ResNext architecture (Xie et al. 2017), we first obtained embeddings of our images. These embeddings represent the image in a high dimensional vector (1,000 in our case), encoding information from the large image datasets they are trained on. We then clustered these embeddings using the k-means clustering algorithm. The number of clusters was decided using the silouhette method (Wang et al. 2017). Examples of image clusters obtained through this method are shown in Figure 10. Each image is a cluster of semantically similar images, representing a semantically meaningful concept. Since this is an unsupervised clustering technique based on a pretrained embedding, the clusters are not always completely coherent. We used this technique to provide additional analysis only on *some* types of images, where the clusters were clean and coherent (e.g. Figure 10).

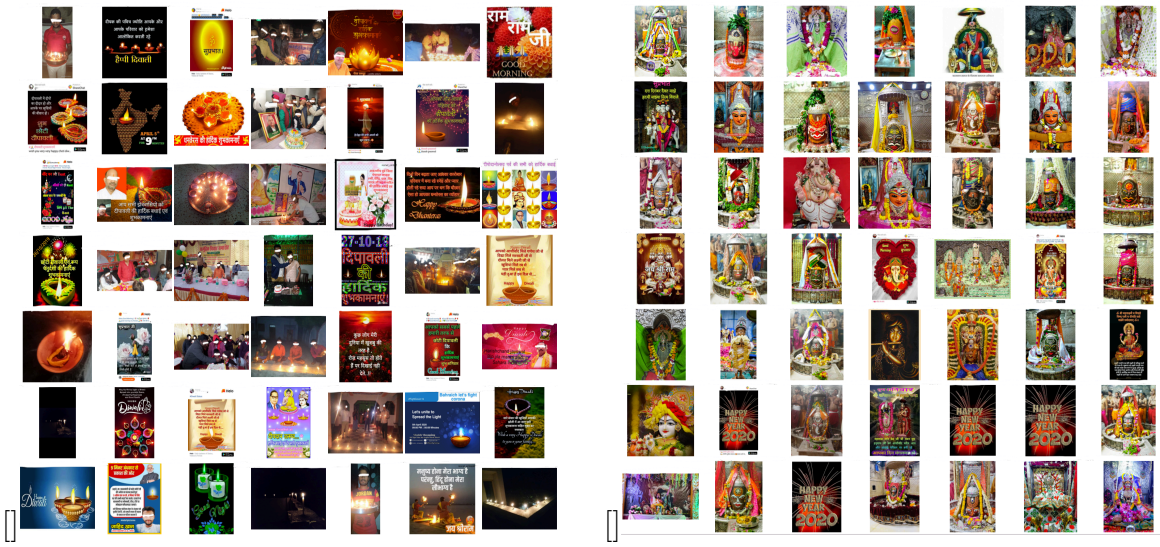

Figure 10. Semantically similar clusters obtained by clustering embeddings obtained from a CNN model: (a) a cluster of images of related to a Hindu festival (Diwali), and (b) a cluster of images of gods.

We obtained 40 clusters using the k-means clustering algorithm. We used the Silhouette method (Wang et al 2017) to obtain the number of clusters. Manually inspecting the clusters, we found that some of them were coherent and correspond to relevant/interesting topics (e.g. pictures of gods).

A qualitative analysis of the clusters reveals that the clusters fall into 5 main themes: 1. political memes, 2. greetings (e.g. good morning, festival wishes, etc), 3. images containing text (including political pamphlets, pictures of news papers, pictures of receipts, etc), 4. images of gods, and 5. gatherings of people.

Fourteen of the forty clusters, though similar to each other semantically, are not interesting descriptively. Examples of two such clusters are shown in Figure 11 below. The remaining 26 clusters mostly include images that are coherent and related to each other. A random sample of 4 clusters which are 'clean' are shown in Figure 12. These four clusters show: (a) Pictures of TV news programs (1,727 images), (b) Good morning messages (3,746 images), (c) Pictures of gods (6,529 images), (d) Pictures of men/women in police uniforms (842 images). These 'clean' clusters can be used to estimate the total number of images belonging to a that category.

There are also large clusters which involve party workers taking selfies and images of gatherings in the dataset, but these are not clean enough to confidently estimate the size of these clusters of images. For instance, Figure 13 shows two such clusters. Figure 13(a) shows a cluster of people gathering (1,589 images), and (b) (mostly) shows people taking selfies (11,840 images). Note that both these clusters do not completely contain only images of people gathering or selfies.

Figure 11. A sample of images from semantically similar clusters of images which are not interesting. (a) images of people wearing sunglasses, (b) memes/images with text on a plain background.

Figure 12. Four examples of 'clean' clusters. (a) Pictures of TV news programs (1727 images), (b) Good morning messages (3746 images), (c) Pictures of gods (6529 images), (d) Pictures of men/women in police uniforms (842 images).

Figure 13. Two clusters of images (a) containing images of gatherings (1589 images) and (b) containing images of gatherings and selfies (11840 images).

APPENDIX C: Additional Analyses / Hateful Content

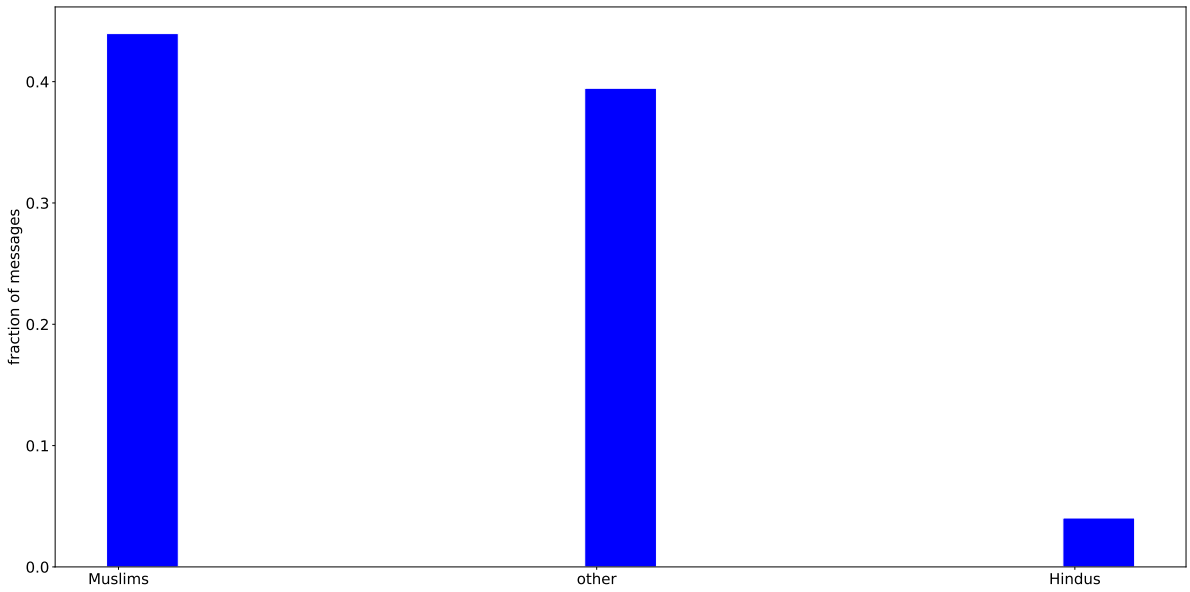

Figure 14. Who is Targeted in Visual Content Labeled as "Hateful"?

Note: the full text for this item was as such: *What is the group targeted?.*

- Muslims
- Hindus
- Christians
- Upper-castes
- Dalits
- Other caste group
- Liberals
- Other: [....]

We only included here responses > 1% of the data.

Figure 15 shows the word cloud of the labels provided by the annotators for the 'Other' group while labeling the group targeted in the hateful content.

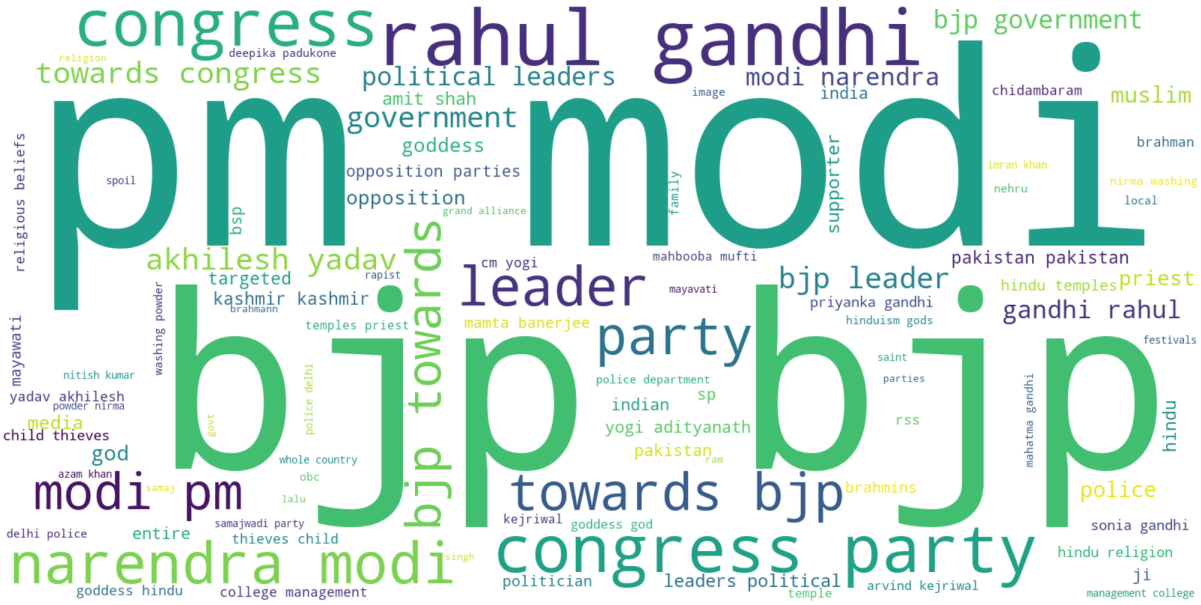

Figure 15. What is the group targeted in the "Other" category? (hateful content)

Automated Detection of Hateful Content

As a strategy to overcome the limitations associated to subjective coding, we additionally make use of high precision key words to identify hateful content. We use a state of the art optical character recognition (OCR), *Tesseract* (Smith et al. 2007), to detect text from the images. *Tesseract* works using an LSTM model which can detect text in over 100 languages. In our case, most of the text is in English or Hindi and can be detected with good accuracy. 35.8% of the images in our entire dataset had at least 10 characters of text detected. Qualitative analysis of the quality of the OCR performance indicated that the quality was close to perfect at detecting words on the image.

Next, we obtained a list of 97 keywords from a prominent Indian NGO working on tracking hate speech on social media.[34] This list consists of high precision swear words used against muslims and other minorities such as "Jihadi", "mullah" etc. We then looked for text

[34]https://www.equalitylabs.org

on the images which mention at least one of these keywords and found that 4,951 images (1.3% of the total images) mention at least one such explicitly hateful keyword. To evaluate the quality of our keyword based approach, we randomly sampled 100 messages containing one of the hate keywords and manually verified whether they were indeed hateful, as per our relatively broad definition of the term. We found that 96% of the messages were correctly classified.

Note that identifying hateful images using keywords mentioned in the text on the image is a crude way of detecting hateful content, and, thus indicates a lower bound of the hate in our dataset. As we saw above, labeling hateful content is hard even for humans, and most hateful content does not necessarily say anything hateful (textually) explicitly. This automated analysis nonetheless echoes the trend mentioned above, showing that explicitly hateful content is quite rare in these WhatsApp groups.

Appendix D: Additional Analyses / Misinformation Content

Much misinformation appears targeted towards an political leaders, as shown in Figure 16.

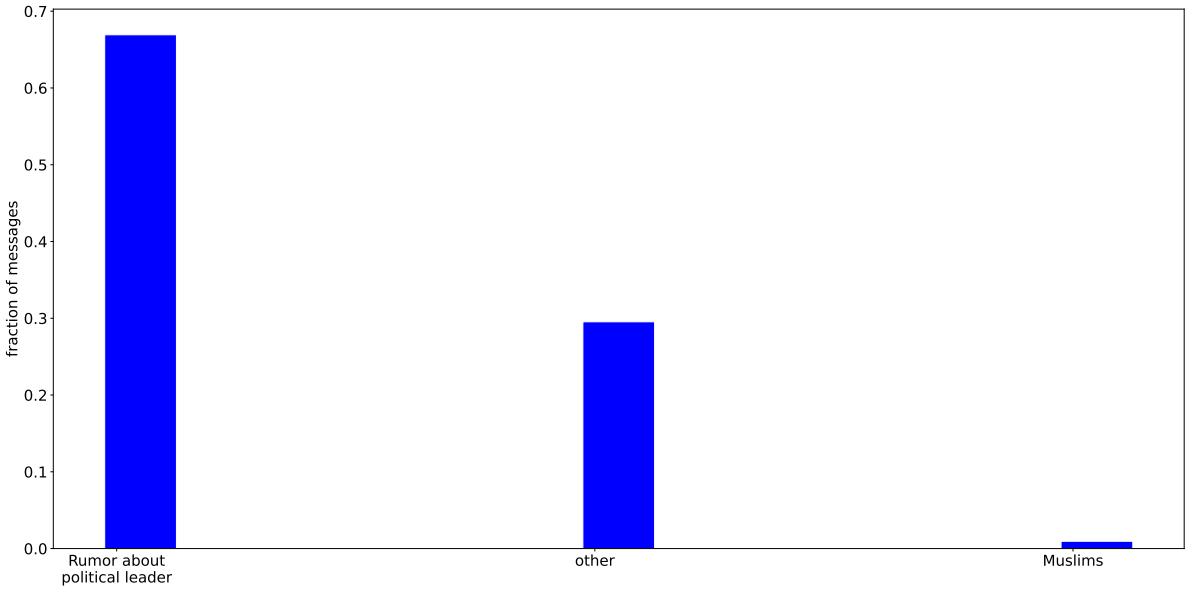

Figure 16. Topic of the Images Coded as containing Misinformation.

Figure 17 shows the word cloud of the labels provided by the annotators for the 'Other' category while labeling misinformation content.

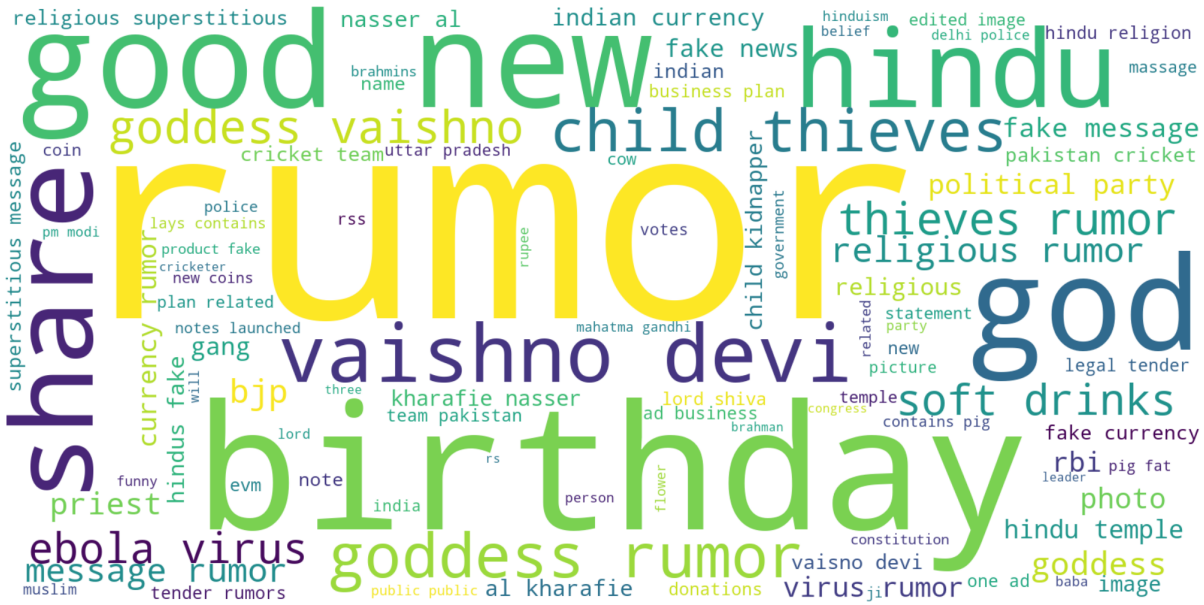

Figure 17. Topic of the "Other" Content - based on coders' descriptions (Misinformation Content).

Automated Detection of Misinformation

To detect content that was misinformation, we also used automated tools built using image hashing techniques mentioned in Appendix B. Using a web scraping script, we obtained all the images from 6 popular fact checking websites in India including AltNews, Boom Live, and SMHoaxSlayer. This gave us 21,000 images. Using perceptual hashing techniques (Garimella et al. 2020), we obtained hashes for these 21k images. Next, we computed the similarity between the hashes obtained on our dataset to the factcheck images to find similar images. The intuition behind this is that if an image that appeared in our dataset was factchecked, it could be misinformation. Overall, we found 4,156 images (1.1% of the total images) which match with at least one fact checking image. While this confirms our above findings, this analysis has certain limitations. The matching using perceptual hashing only works if the images have *not* been altered or cropped. Also, we assume here that all the images which match with an image from a fact checking website is misinformation. This might not always

be the case. Nonetheless, we take this as yet additional evidence of our main finding so far, that misinformation remains overall relatively rare on these threads.

APPENDIX E: Analyses of Tags

Given the limitations associated to relying on subjective codings, we also evaluate the extent to which the content of partisan threads focuses on topics that we know to be, from the recent literature on India, associated with some heated topics. Accordingly we calculate here the fraction of posts classified by at least one of the two coders as being associated to each of the following tags, both for the full sample and for the BJP subsample: *Violence/Muslims/Hinduism/Crime/Kashmir/NRC-CAB/Economic performance OR achievements of the government/National security/Pakistan/Political leaders/Political leaders (national)/Political leaders (local)/Wishes and greetings/Fun OR timepass OR entertainment/None of the above.*

In keeping with previous findings, much of the content appears to be about political leaders, even if other topics are tagged as well, to a lower extent. Importantly, around 70% of the content tagged as being about political leaders is additionally about *local* leaders (as opposed to National leaders).

Figure 18 contains the tags associated with the images, in the complete dataset.

Figure 19 shows the same figure for a subset of only BJP groups.

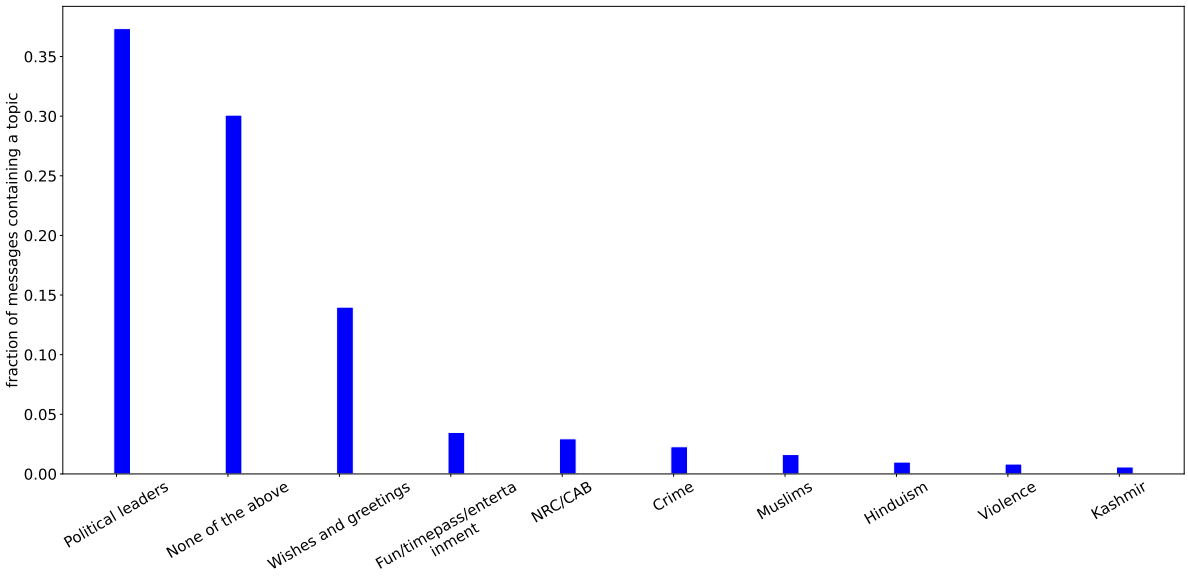

Figure 18. Tags associated to coded images.
(full dataset)

Note: the full text for this item was as such:

Is this content about any of the following topics? (TICK AS MANY AS APPLY)

- Violence
- Muslims
- Hinduism
- Crime
- Kashmir
- NRC/CAB
- Economic performance/achievements of the government
- National security

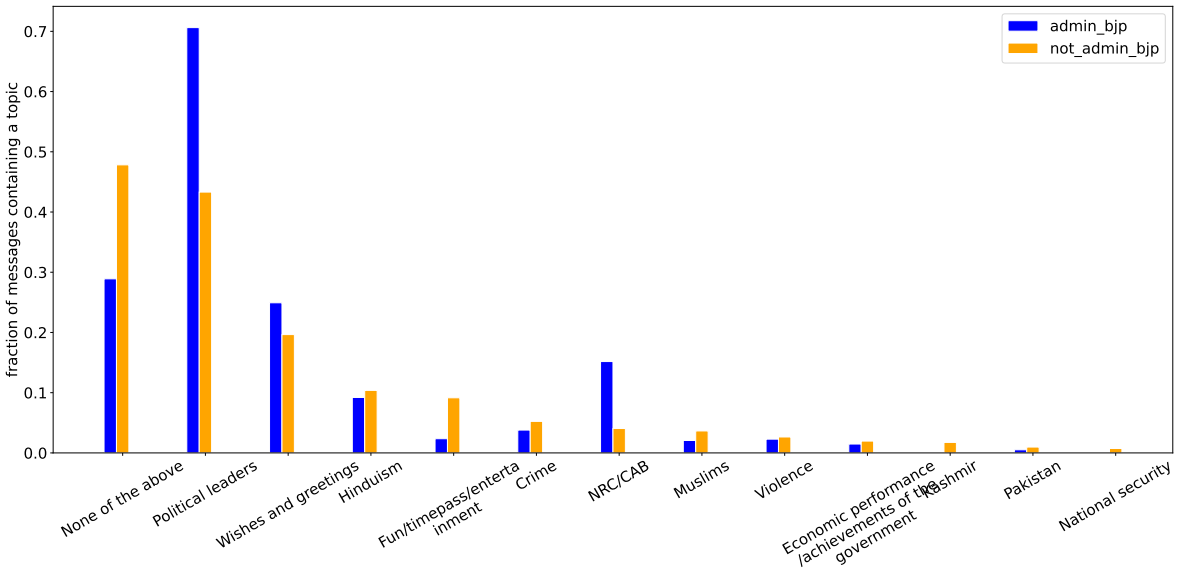

Figure 19. Tags Associated to Coded Images (BJP groups subsample - broken down by group admins vs. other users).

- Pakistan
- Political leaders
  - Political leaders (national)
  - Political leaders (local)
- Wishes and greetings
- Fun/timepass/entertainment
- None of the above

We only included here responses > 1% of the data.

Appendix F: Additional Analyses / Partisan Content

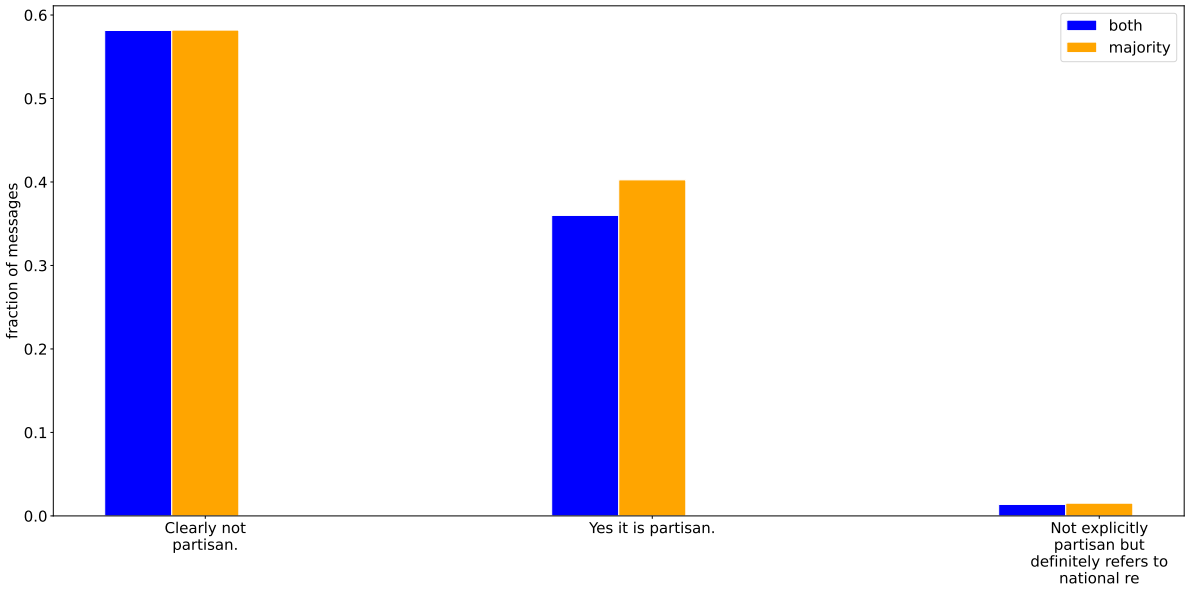

Figure 20: Comparing amount of "partisan content" based on agreement between initial coders (yellow) vs. subset in which disagreements were adjudicated by a third coder and included if 2/3 coders counted it as partisan (blue)

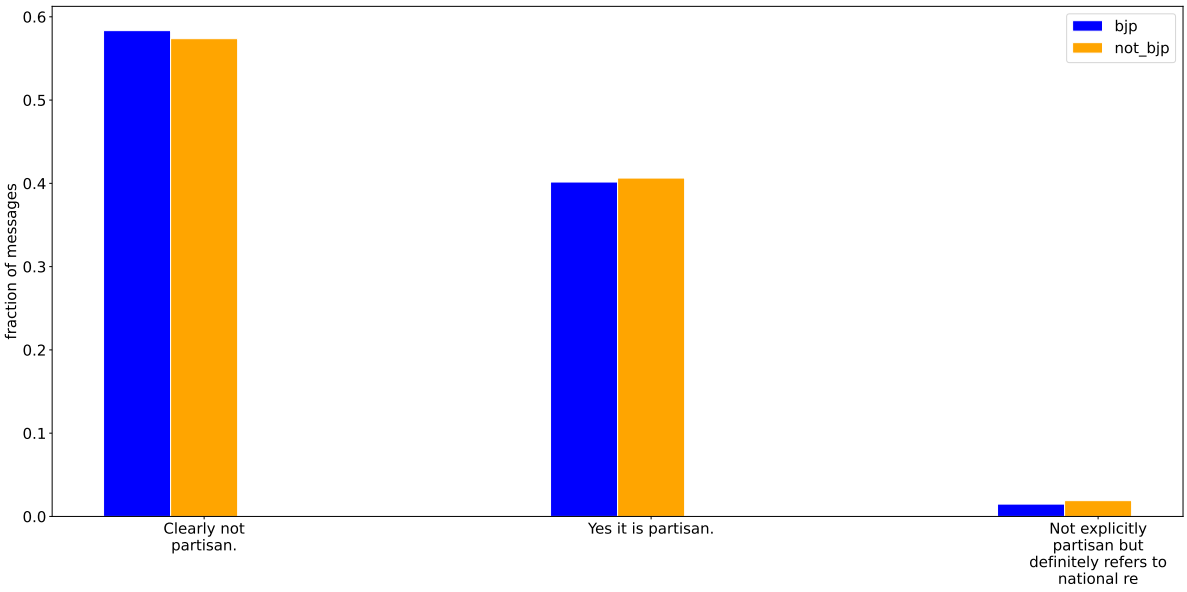

Figure 21. Partisan Content on BJP vs non BJP threads."

Figure 22 shows the reason given by the coders for labeling an image as partisan.

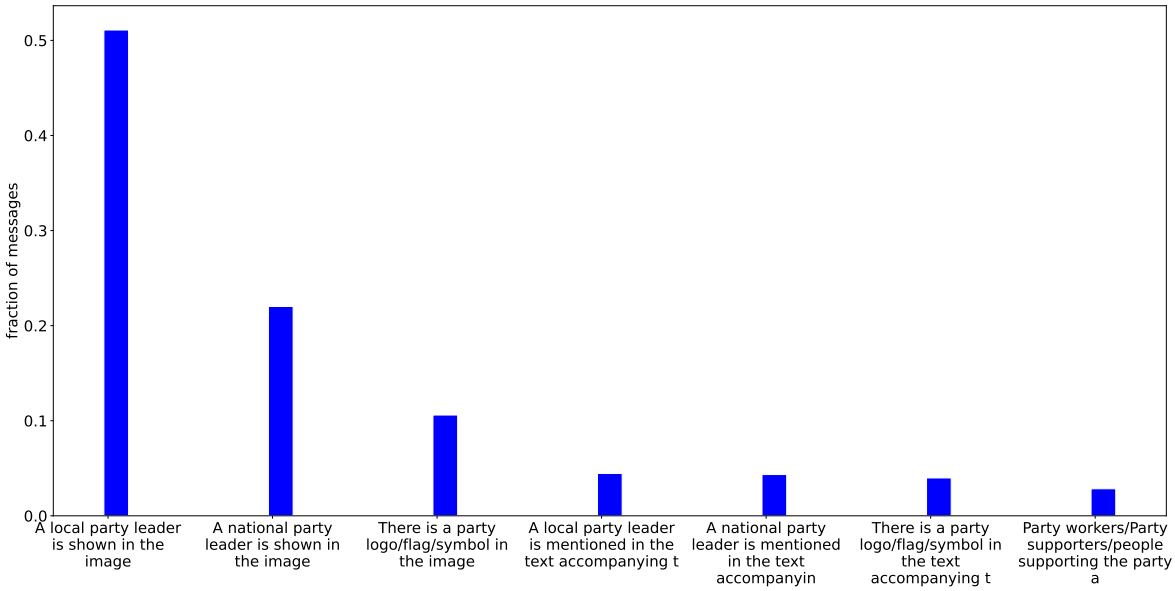

Figure 22. Rationale given by coders for coding images as "partisan content."

Note: the full text for this item was as such: *Why did you code this as "partisan"? (go through the list and check **ALL that apply**; check at least one; fill in comment box if need be).*

- a party message is included ON the image (as text)
- a party message is included in the text accompanying the image (either in the same message or in directly preceding/following message posted by same individual)
- A national party leader is shown in the image
  - PM Modi is shown in the image
  - Rahul Gandhi is shown in the image
- A local party leader is shown in the image
- Party workers/Party supporters/people supporting the party are shown in the image

- A national party leader is mentioned in the text accompanying the image
  - PM Modi is mentioned in the text accompanying the image
  - Rahul Gandhi is mentioned in the text accompanying the image
- A local party leader is mentioned in the text accompanying the image
- There is a party logo/flag/symbol in the image
- There is a party logo/flag/symbol in the text accompanying the image item
- The image/video generally implies a partisan message or conveys one or several ideas congruent to the party's ideology (or critical of an opposition party).

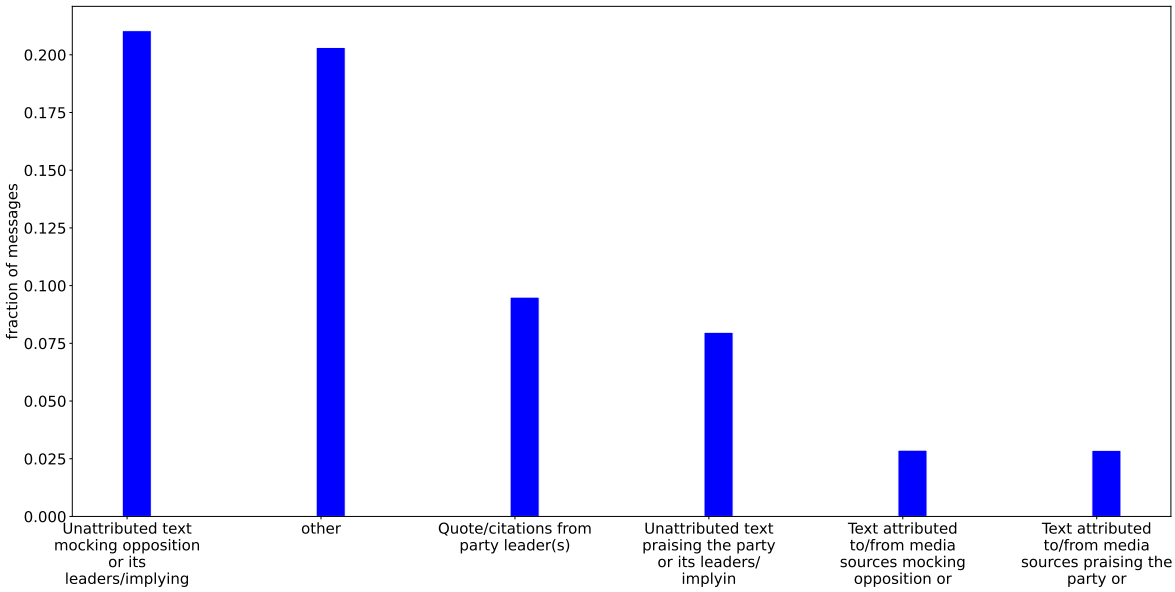

Figure 23. If the image contains "partisan text", what type of partisan text does it include?

Note: the full text for this item was as such:

What kind of partisan text does the image contain? (check ALL that apply; check at least one).

- Quote/citations from party leader(s). .
- Unattributed text praising the party or its leaders/ implying that the party or its leaders did something good (i.e. positive news).
- Unattributed text mocking opposition or its leaders/implying that the opposition or its leaders did something wrong (i.e. negative news).
- Text attributed to/from media sources praising the party or its leaders/ implying that the party or its leaders did something good (i.e. positive news).
- Text attributed to/from media sources mocking opposition or its leaders/ implying that the opposition or its leaders did something wrong (i.e. negative news).

- Other [...].

We only included here responses > 1% of the data.

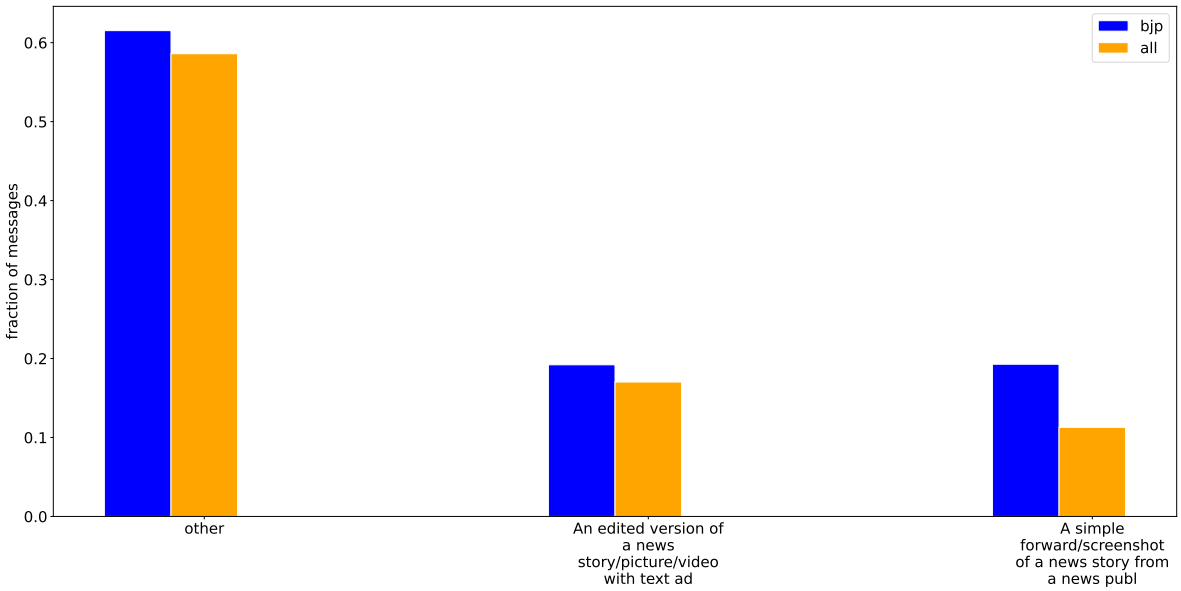

Figure 24. Is this partisan image...?

Note: the full text for this item was as such: *is the image….?*

- A simple forward/screenshot of a news story from a news publication/channel.
- An edited version of a news story/picture/video with text added on top etc (ex: a meme), the source of which is not clear.
- An edited version of a news story/picture/video with text added on top etc (ex: a meme), the source of which appears to be partisan (ex: BJP IT cell; BJP4India etc).
- Other [...]

We only included here responses > 1% of the data.

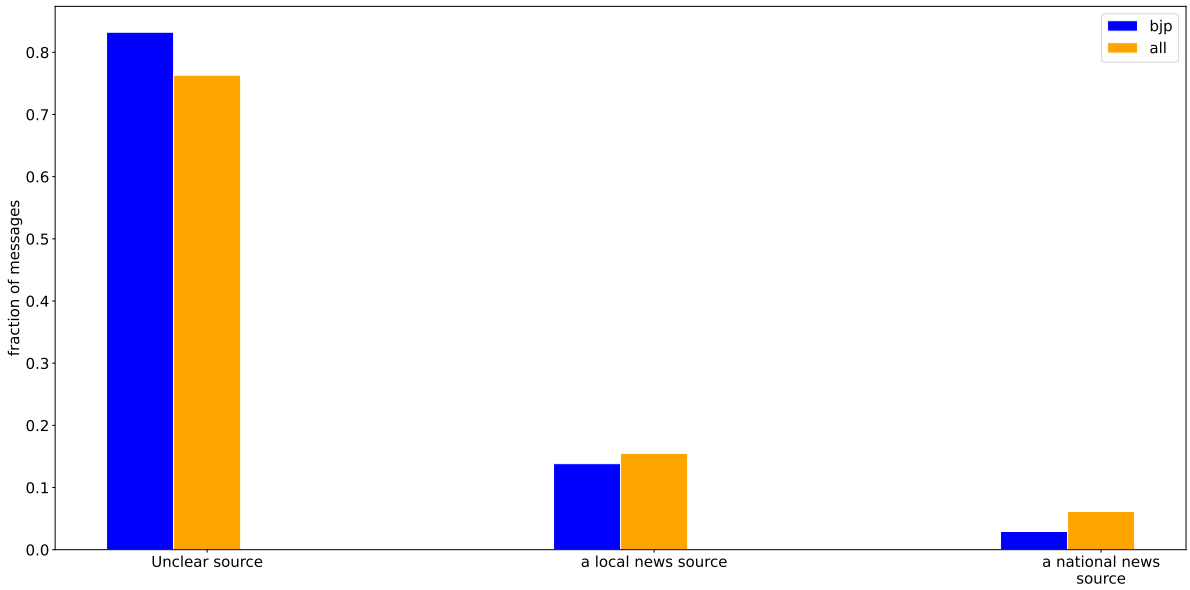

Figure 25. Source of the image

Note: the full text for this item was as such: *Is this from ….*

- a local news source
- a national news source
- unclear source

We only included here responses > 1% of the data.

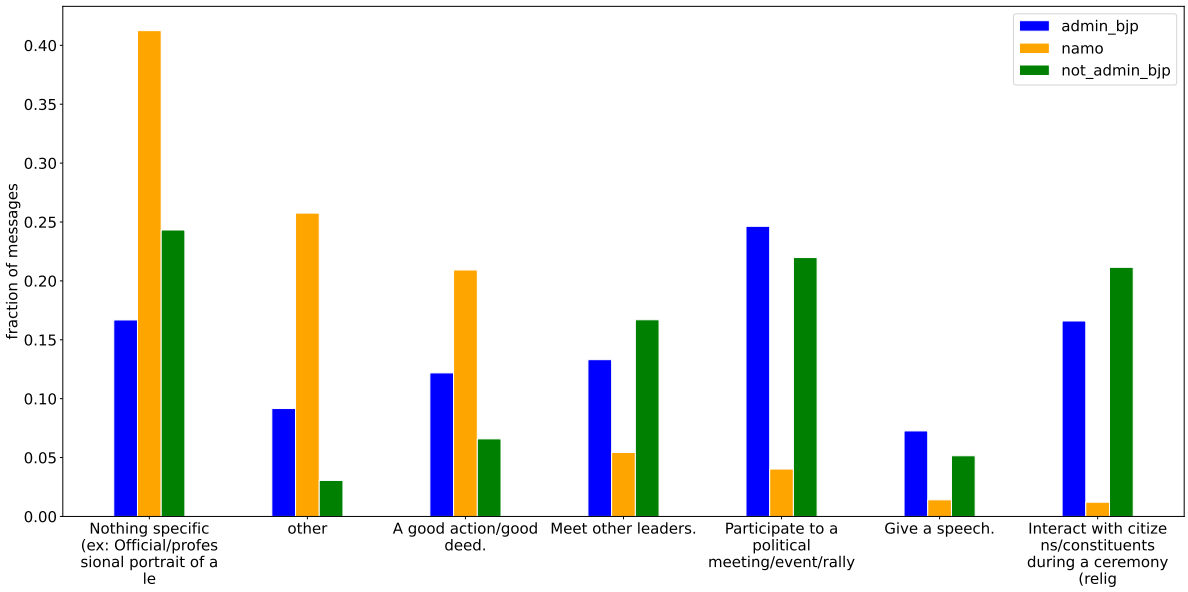

Figure 26. If the image contains party leaders or members, what are they doing?

Note: the full text for this item was as such:

You noted that the image/video contains party leaders or members. What are they doing on the picture/video? (check ALL that apply; check at least one).

- A good action/good deed.
- Interact with citizens/constituents during a ceremony (religious or official or political).
- Meet other leaders.
- Give a speech.
- Participate to a political meeting/event/rally
- Something else.
- Nothing specific (ex: Official/professional portrait of a leader)
- Other [...]

We only included here responses $>$ 1% of the data.

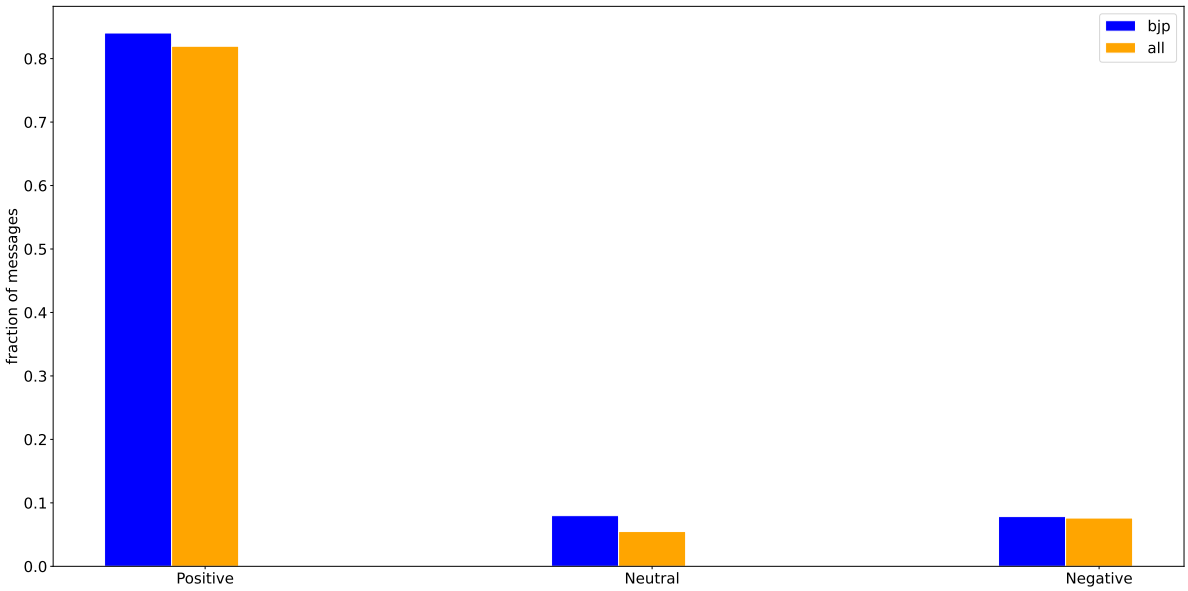

Figure 27. Is the partisan content positive, negative or neutral?

APPENDIX G: Sampling / Descriptive Statistics on the "Social Media Workers" whose Partisan Groups are Included in the Sample.

	Overall sample of social media workers who were asked about the thread they maintain **(N=1547)**	Among those who gave the name of the group (**87% of the total)**	Among those who verbally agreed to add us to the sample (**52% of the total**)	Among those whose groups we could actually study (**34% of the total**)
% male	99.16	99.11	99.38	99.25
Mean age	35.3	35.05	34.39	**33.86**
Mean years of education	11.94	12.00	12.02	12.00
Mean income (monthly, in rupees)	15,194	15,410	16,961	18162
% who declare following political news and events "on a daily basis".	67.23	67.54	66.83	66.1
% BJP	68.03	69.85	71.53	70.24
% in BJP subset belonging to "IT Cell"	7.03	7.26	7.27	7.24
% in BJP subset belonging to "RSS or VHP"	11.98	12.5	11.07	10.99
% Currently holding a formal position in their party	81.76	82.59	82.51	84.31
Mean number of years supporting the party	10.62	10.69	10.84	**9.45**
Mean response to: *"On how many days each month do you send messages on behalf of the party on social media?"*	16.03	16.42	16.26	16.28
Mean response to: *"Looking at your phone, how many {PartyName} related threads would you say you are on, approximately?"*	8.81	9.62	9.53	9.31
Mean response to: *"Approximately how many of these do you actually post messages on?"*	5.57	6.08	5.75	5.73
Mean response to: *"How many of these do you forward materials from to other threads?"*	4.72	5.17	4.68	4.69
Mean response to: *"How many of these local party threads are you yourself the admin for?"*	1.02	1.14	1.25	**1.31**

Figure 28. Characteristics of Re-contacted Social Media Workers

APPENDIX H: Characteristics of Coders

In Table 29, we detail the characteristics of the 12 human coders that participated in the coding of the materials described in the core of the article.

Median age	27.5	(N/A)
% male	75.00%	(9/12)
% with advanced university degree (superior to BA or BS)	66.67%	(8/12)
% with advanced level (or above) in English	75.00%	(9/12)
% fluent (native level) Hindi	100.00%	(12/12)
% Grew up in UP	100.00%	(12/12)
% Hindu	100.00%	(12/12)
% Upper Caste	50.00%	(6/12)
% Lower Caste	33.33%	(4/12)
% Dalits and Muslims	16.67%	(2/12)
% Supporting Modi Government	50.00%	(6/12)
% Opposing Modi Government	50.00%	(6/12)
% Supporting (or strongly supporting) BJP State Government	41.67%	(5/12)
% Opposing (or strongly opposing) BJP State Government	58.33%	(7/12)
% Reported being close or very close to BJP	41.67%	(5/12)
% Reported being close or very close to INC	0.00%	(0/12)
% Reported being close or very close to BSP	33.33%	(4/12)
% Reported being close or very close to SP	41.67%	(5/12)
% Using WhatsApp several times/day or more	91.67%	(11/12)
% Reported being on a partisan WhatsApp thread on their private phone	50.00%	(6/12)
% Declared following news about UP politics closely or very closely	100.00%	(12/12)
% Declared following news about National politics closely or very closely	91.67%	(11/12)
% Declared watching Hindi news channels daily	58.33%	(7/12)
% Declared watching English-language news channels daily	33.33%	(4/12)
% Declared reading a Hindi newspaper daily	50.00%	(6/12)
% Declared reading an English-language newspaper daily	8.33%	(1/12)

APPENDIX I: Additional Analyses / "Other" Category in Figure 4

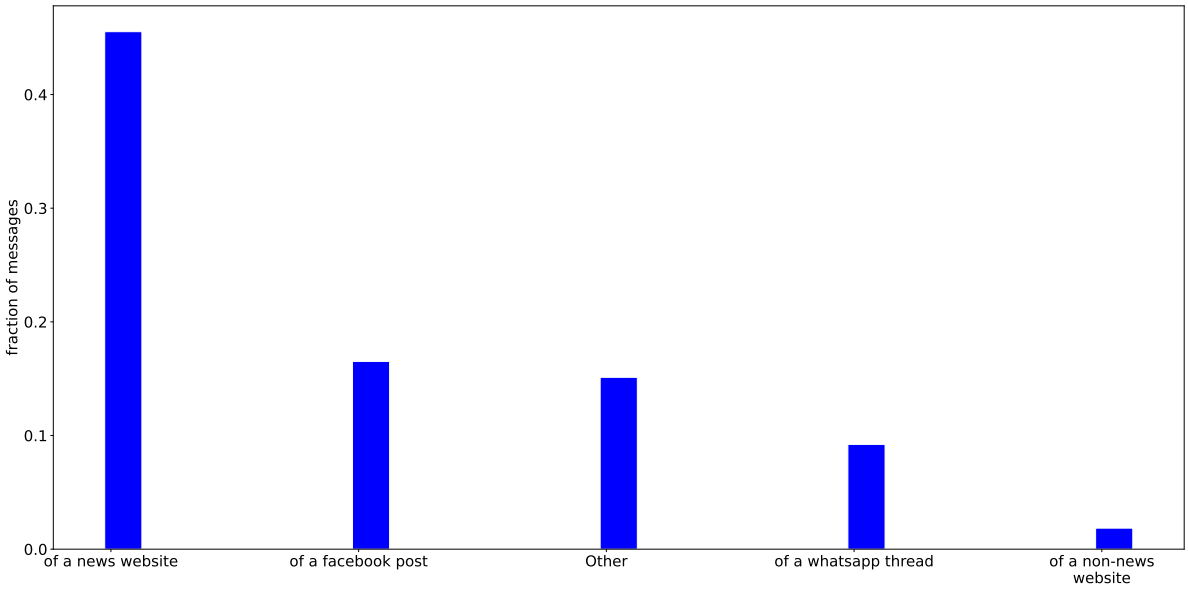

Figure 30. If a screenshot, what type of screenshot is it?

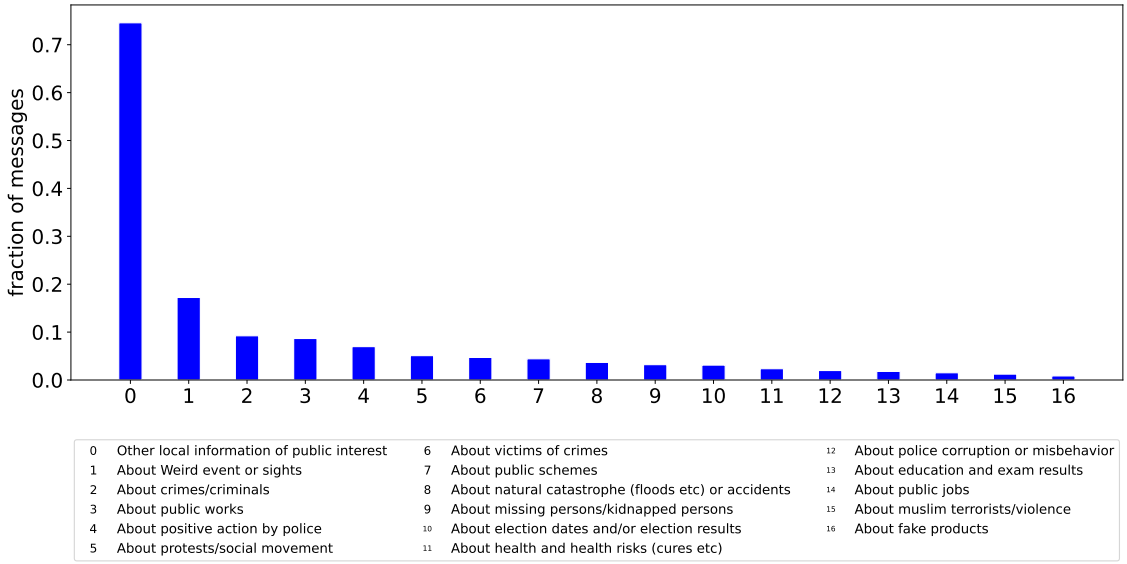

Figure 31. If a screenshot or an image of a newspaper article, what is the content of the screenshot or article about?

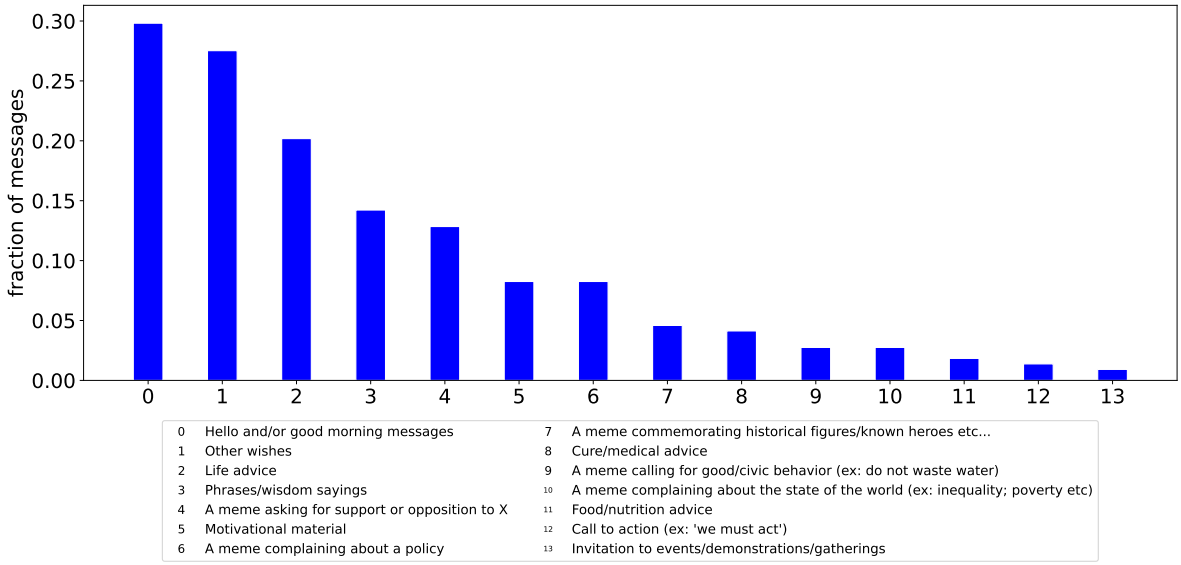

Figure 32. If a meme, what type of meme is it?

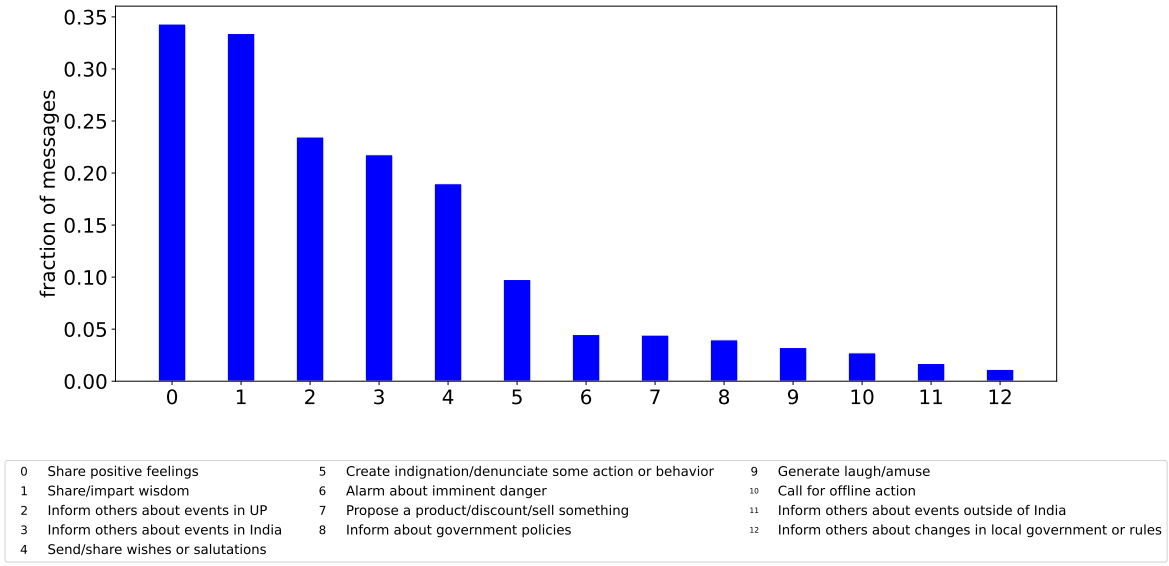

Figure 33. What do you think the main objective of the posted image is?