Persistence of online misinformation: Investigating Facebook's actions against "repeat offenders"

Anonymous TTO submission

Abstract

Like most web platforms, Facebook is under pressure to regulate misinformation. According to the company, pages that repeatedly share misinformation ("repeat offenders") will have their distribution reduced, but little is known about the implementation or the efficacy of this measure. We aimed to investigate the implementation and consequences of this policy using a first of its kind analysis, combining data from a fact-checking organization, users' self-declaration and CrowdTangle data. We did not observe that accounts repeatedly sharing misinformation had reduced engagement metrics, but a drastic 50% drop was observed around June 9, 2020. No public information was given by Facebook about this sudden decrease. Overall, we find no evidence so far that Facebook's reduced distribution policy against repeat offenders is having any impact on misinformation distribution.

1 General introduction

With an ever-increasing proportion of the public getting their information online, mainly through search engines, social media and video platforms (Mitchell et al., 2016), the spread of misinformation through these platforms has received growing attention. Recent studies and the political context of January 2021 show how the presence of misinformation online can contribute to negative societal consequences by fueling false beliefs, such as the idea that massive voter fraud occurred during the US 2020 presidential election, which contributed to the January 6, 2021 insurrection at the U.S. Capitol (Benkler et al., 2020) and other false stories about presidential candidates (Allcott and Gentzkow, 2017). Misinformation has also confused the public about the reality of climate change (Brulle, 2018; Porter et al., 2019) and stoked skepticism about vaccine safety among the public (Featherstone and Zhang, 2020; Lahouati et al., 2020). In April 2020, a questionnaire from the Reuters Institute found that people in the UK use online sources more often than offline sources when looking for information about the coronavirus. Among social media platforms, Facebook was the most widely used with 24% of the respondents saying they used Facebook to access COVID-19 information in the last seven days (Fletcher et al., 2020). The structural importance of Facebook to the media landscape is confirmed by Parse.ly's dashboard, showing that the visitors to their 2500+ online media sites are referred by Facebook in 25% of the cases, second to Google's referral volume accounting for 54% of traffic[1].

Lawmakers and regulators are increasingly pressuring platforms to limit the spread of disinformation. In the US, the House of Representatives organized hearings and convoked representatives of the main platforms to shed light on how they are being weaponized to spread "misinformation and conspiracy theories online" (Donovan et al., 2020). In Europe, the European Commission has established a 'Code of Practice on Disinformation'[2] that enjoins platforms to voluntarily comply with a set of commitments (Heldt, 2019). However, there is little data available and few established processes to monitor the implementation of these measures and quantify their actual impact. This is what we propose to tackle in this paper by offering a methodology to monitor Facebook's implementation of one of its core policies against misinformation. We chose to focus on Facebook as it is the biggest social media platform with more than 2 billion users worldwide.

Facebook announced a three-part policy to fight

[1] https://www.parse.ly/resources/data-studies/referrer-dashboard, accessed on 2021-07-08.

[2] https://ec.europa.eu/digital-single-market/en/code-practice-disinformation.

against 'misleading or harmful content': they claim to *remove* harmful information, *reduce* the spread of misinformation and *inform* people with additional context[3]. Facebook has developed the most extensive third-party fact-checking program with dozens of partner instutition to assist the company in this endeavour[4]. When a fact-checking partner flags a URL, a post or a video as misinformation, Facebook claims to display the posts marked as "False" or "Partly False" further down in users' feed, further reducing the virality of these posts. Facebook also informs page or group owners when published posts on pages or groups that they manage are marked as misinformation, inviting them to correct the posts. Facebook's *reduce* policy is not only applied to individual posts, but also to organizations and communities that often publish posts containing misinformation, as indicated by this statement in their publishers' help center[5]:

> *Pages and websites that repeatedly share misinformation rated False or Altered will have some restrictions, including having their distribution reduced.*

So far Facebook has not provided data showing how their reduce policy is implemented, which would allow researchers to quantify its impact on misinformation circulation. To the best of our knowledge, the impact of the reduce policy has not yet been audited directly. It is in this way that the present research paper distinguishes itself from the articles that measured overall levels of misinformation on the platform (Allcott et al., 2019; Kornbluh et al., 2020; Resnick et al., 2018).

CrowdTangle, a public insights tool owned and operated by Facebook, was used to access Facebook data (Team, 2021). CrowdTangle exclusively tracks public content, and provides access to engagement metrics (such as number of likes, shares and comments), but not to the reach (number of views) of content[6]. We first investigated how Facebook enforces its 'reduce' policy by combining data from a Facebook fact-checking partner identifying URLs sharing misinformation and tracking engagement metrics of the Facebook accounts that repeatedly share such misinformation. We then further investigated the effects of Facebook's policy on engagement metrics of a set of Facebook pages claiming to be under reduced distribution.

2 Investigating the 'reduce' policy on Facebook groups repeatedly sharing misinformation

To investigate the effect of fact-checking on Facebook accounts that repeatedly share misinformation, we first used data from Science Feedback, which is part of Facebook's third-party fact-checking program[7].

2.1 Methods

Science Feedback is an fact-checking organization, in which academics are verifying the credibility of science-related claims and articles. Out of the 4,000+ URLs labeled by Science Feedback, we relied on the 2,452 URLs labeled as 'False', which we call "false news links". Were excluded the URLs labeled as 'Partly False', 'Missing Context', 'False headlines' or 'True', as well as the URLs marked as 'False' but 'corrected to True' by the publisher, since these labels do not contribute to the repeat offender status according to Facebook's guidelines. The list of false news links was obtained on January 4, 2021 and cover links flagged in 2019 and 2020.

Using the '/links' endpoint from the CrowdTangle API, we gathered the public Facebook groups and pages that shared at least one false news link between January 1, 2019 and December 31, 2020. Due to the API limitations, if a URL was shared in more than 1000 posts, we collected only the 1000 posts that received the highest number of interactions[8]. We focused on the accounts that spread the most misinformation, and chose a threshold of 24 different false news links shared over the past two years.

The corresponding 307 Facebook accounts (289 Facebook groups and 18 Facebook pages) are named 'repeat offenders accounts'. All the posts they published between January 1, 2019 and December 31, 2020 were collected using the '/posts'

[3] https://about.fb.com/news/2018/05/inside-feed-reduce-remove-inform/

[4] https://www.facebook.com/business/help/341102040382165

[5] https://www.facebook.com/business/help/2593586717571940, https://www.facebook.com/business/help/297022994952764

[6] https://help.crowdtangle.com/en/articles/3192685-citing-crowdtangle-data, https://help.crowdtangle.com/en/articles/4558716-understanding-and-citing-crowdtangle-data

[7] https://sciencefeedback.co/science-feedback-partnering-with-facebook-in-fight-against-misinformation/

[8] https://github.com/CrowdTangle/API/wiki/Links

endpoint. We gathered the daily number of posts, and calculated the total and average daily number of interactions per post, aggregating the comments, shares and reactions (such as 'like', 'love', 'favorite', 'haha', 'wow', 'sad' and 'angry' reactions).

'Repeat offenders' accounts are supposed to have their distribution reduced, according to Facebook's official communication, but the precise rule Facebook uses to classify an account as 'repeat offenders' is not specified. An undisclosed source obtained by a journalist indicated that "The company operates on a 'strike' basis, meaning a page can post inaccurate information and receive a one-strike warning before the platform takes action. Two strikes in 90 days places an account into 'repeat offender' status"[9].

Based on this 'two strikes in 90 days' rule and the list of strike dates known by Science Feedback, we inferred periods during which each account must have been under repeat offender status. If a post sharing a misinformation link was published after the corresponding fact-check, we used the date of the post as the strike date. If the account first shared a link, which was then fact-checked as 'False', the fact-check publication date was used as the strike date. Any given time in which an account shared two or more false news links over the past 90 days is defined as a repeat offender period.

2.2 Results

Figure 1's upper panel displays the engagement metrics for one 'repeat offender' example group named 'Australian Climate Sceptics Group'. The known strike dates are shown as red lines at the bottom, and the inferred 'repeat offender' periods are shaded in red. The daily engagement metrics (number of reactions, shares and comments per post averaged per day) vary throughout the measuring period, but they do not appear to be related with the alternance of 'repeat offender' and 'no strike' periods.

[9] https://www.nbcnews.com/tech/tech-news/sensitive-claims-bias-facebook-relaxed-misinformation-rules-conservative-pages-n1236182

3 Investigating the 'reduce' policy on self-declared 'repeat offenders' Facebook pages

4 General discussion

References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36.

Hunt Allcott, Matthew Gentzkow, and Chuan Yu. 2019. Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2):2053168019848554.

Yochai Benkler, Casey Tilton, Bruce Etling, Hal Roberts, Justin Clark, Robert Faris, Jonas Kaiser, and Carolyn Schmitt. 2020. Mail-in voter fraud: Anatomy of a disinformation campaign. *Available at SSRN*.

R Brulle. 2018. 30 years ago global warming became front-page news–and both republicans and democrats took it seriously. *The Conversation*.

J. Donovan, N. Jankowicz, C. Otis, and M. Smith. 2020. House intelligence committee open virtual hearing: "misinformation, conspiracy theories, and 'infodemics': Stopping the spread online". https://intelligence.house.gov/news/documentsingle.aspx?DocumentID=1092.

Jieyu Ding Featherstone and Jingwen Zhang. 2020. Feeling angry: the effects of vaccine misinformation and refutational messages on negative emotions and vaccination attitude. *Journal of Health Communication*, 25(9):692–702.

Richard Fletcher, Antonis Kalogeropoulos, Felix M Simon, and Rasmus Kleis Nielsen. 2020. Information inequality in the uk coronavirus communications crisis. *Reuters Institute for the Study of Journalism*.

Amélie Heldt. 2019. Let's meet halfway: Sharing new responsibilities in a digital age. *Journal of Information Policy*, 9:336–369.

K Kornbluh, A Goldstein, and E Weiner. 2020. New study by digital new deal finds engagement with deceptive outlets higher on facebook today than run-up to 2016 election. gmf the german marshall fund of the united states. viitattu 16.12. 2020.

Marin Lahouati, Antoine De Coucy, Jean Sarlangue, and Charles Cazanave. 2020. Spread of vaccine hesitancy in france: What about youtube™? *Vaccine*, 38(36):5779–5782.

Amy Mitchell, Jeffrey Gottfried, Michael Barthel, and Elisa Shearer. 2016. The modern news consumer: News attitudes and practices in the digital era. *Pew Research Center*.

Ethan Porter, Thomas J Wood, and Babak Bahador. 2019. Can presidential misinformation on climate change be corrected? evidence from internet and phone experiments. *Research & Politics*, 6(3):2053168019864784.

Paul Resnick, Aviv Ovadya, and Garlin Gilchrist. 2018. Iffy quotient: A platform health metric for misinformation. *Center for Social Media Responsibility*, 17.

CrowdTangle Team. 2021. Crowdtangle. *Facebook, Menlo Park, California, United States*.