Very Preliminary Draft

June 30, 2021

Using social media data to verify the platforms' regulation policies regarding "misinformation"

1 Introduction

A number of recent studies point towards the idea that "Fake News" or disinformation is a small subset of the total supply of information on online social networking platforms (e.g. Grinberg et al. (2019) [2] and Broniatowski et al. (2020) [1]). Yet, this seemingly small subset is generating great concern in traditional media and in society in a broader sense. [1]

Section 230 in the United States Communications Decency Act[2] provides immunity for website platforms against the content created by users. Nevertheless, there is growing pressure for Mainstream Social Media Platforms (hereafter MSMP), such as Facebook, Twitter or Youtube, to moderate the available content. In particular, platforms seem to take explicit actions when content is in violation of local laws in different jurisdictions, e.g. laws regarding defamation of a racial nature, dissemination of symbols from unconstitutional organizations, privacy protection, digital security, electoral laws. Facebook reports having implemented a total of 64.7 thousand content restrictions based on local law across all countries in 2020.[3] Google reports a total of 26 thousand government requests to remove content from July 2020 to December 2020, among which 11.4 thousand concerned Youtube.[4] Twitter reports having received 42.2 thousand legal demands from third-parties from January to June 2020, and has responded by withholding 82 thousand accounts and 3.1 thousand tweets.[5]

Furthermore, MSMP are increasingly engaging in editorial tasks by implementing targeted policies to insure that each platform's rules are not violated. Community guidelines of Facebook, Twitter and Youtube can be summarized in a handful of categories, regarding safety, privacy and authenticity; which include violence, terrorism, child sexual exploitation, abuse, harassment, hateful conduct, suicide or self-harm, illegal or regulated goods and services, platform manipulation and

[1] For example see the February 2020 speech of the Director General of the WHO at the Munich Security Conference, where he says "But we're not just fighting an epidemic; we're fighting an infodemic."

[2] Similar regulation exists in the European Union, see articles 12 and 15 of the E-commerce Directive (2000).

[3] See Facebook Transparency Center, Content restrictions based on Local Law: transparency.fb.com/data/contentrestrictions. We summed the count of content restrictions over all countries reported in the table, for $H1$ and $H2$ of the year 2020.

[4] See Google's Transparency report, government requests to remove content: transparencyreport.google.com/government-removals/overview.

[5] See Twitter Transparency website, Removal requests: transparency.twitter.com/en/reports/removal-requests.html#2020-jan-jun.

spam.[6] While specific to each platform, the previously cited categories correspond in most cases to well defined concepts that fall into legal frameworks in many countries.

In this article, we focus on MSMP's policies and actions regarding content with low credibility or false information, commonly referred to as *Fake News*.[7] The *Fake News* phenomenon is still ill-defined by the academic community as it encompasses several combined features such as spreading inaccurate, false or misleading information, with or without the intention of influencing or manipulating a target pool of audience. The rise of social networking platforms over the last decade in terms of number of users worldwide and volume of content, has modified the information ecosystem in terms of production of information and its mediation. Many users can now produce and share content which includes news related information, without having to abide by strict editorial processes that ensure accuracy of information and reliability of sources. Fact-checkers + no laws or not so much.

During the COVID19 global health pandemic platforms have upgraded their guidelines to include a set of rules to tackle the propagation of potentially harmful content.[8] As each platform is a private company, those *new* policies are not coordinated and are implemented in different ways across platforms. Such targeted policies show the willingness of MSMP to enhance the quality of the online conversation, but also sheds light on the lack of specific policies to tackle misinformation in general. Say what they say they do + partnerships with fact-checkers + algorithms + community, Make point about "recycling" existing policies (e.g. to tackle terrorism) and apply it to misinformation+ idea about academic community needs to be able to study this phenomenon and assess the impact of the policies, what effects they have etc.

In the present article, we will explain how to verify with data mining MSMP's actions regarding content with low credibility or false information, through a series of examples for different actions and platforms. For the purpose of clarity, we only focus on three platforms: Facebook, Twitter and Youtube. Both Facebook and Youtube are in the top three most popular social media platforms in terms of number of users.[9] We further choose Twitter because it is a social networking platform with the most news-focused users, according to the Pew Research center (2019) [3]. More specifically, we survey a number of common policies used in order to tackle misinformation, across the three above cited MSMP, Facebook, Twitter and Youtube: temporary or permanent suspension of users, reducing the visibility of some content, introducing flags and notices. We do not provide an exhaustive list of methods on how to investigate the platforms' policies. We rather provide a methodology to investigate key policies, that can be useful to researchers or journalists interested in implementing external monitoring. We summarize in table 1 the tools used to collect the data from Facebook, Twitter and Youtube, that we use in multiple examples throughout the present article. Finally, we discuss how an increased effort of transparency regarding specific content can help the community of researchers study and assess the impact of platforms' policies regarding misinformation.

[6] For an exhaustive overview of the community standards of Facebook, see: facebook.com/communitystandards/. For the Twitter Rules see: help.twitter.com/en/rules-and-policies/twitter-rules, and for Youtube community guidelines see: youtube.com/intl/en_us/howyoutubeworks/policies/communityguidelines/.

[7] For an overview on the *concept* of *Fake News*, we refer the reader to the article The science of fake news by Lazer et al. (2018) [4].

[8] For Facebook, Twitter and Youtube see respectively the following updates:

[9] See for example the ranking of the most popular social networks as of April 2021 on Statista: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

	Application Programming Interface (API)	Web Scraping
Facebook	CrowdTangle: public content insights tool owned and operated by Facebook. Buzzsumo: commercial content database	
Twitter	Twitter API V2	Minet twitter scrape
Youtube	Youtube API V3	

Table 1: Data collection

2 Policies

2.1 Temporary suspension and Permanent suspension

Main stream social media platforms such as Facebook, Twitter and Youtube, may suspend the account of a specific user when they deem that the platforms' rules have been violated. Account suspension can be temporary or permanent.[10] When the suspension is temporary the user is prohibited for a limited period of time from posting content on their account, but created content prior to suspension remains available to the user and their followers. However, when the suspension is permanent, in most cases, followers or subscribers have no longer access to the content prior to the suspension and the user can no longer use the account to create new content. In what follows, we focus on the implementation of this policy by several platforms and provide simple examples to illustrate.

2.1.1 Facebook

When an account is permanently suspended by Facebook, it disappears from the platform (so its data cannot be scrapped anymore), and its data also disappears from the CrowdTangle API.[11]

Interestingly, Facebook is regularly publishing a monthly *coordinated inauthentic behavior* report where it informs how many personal accounts, pages or groups were deleted and to which *deceptive network* they may have belonged.[12] But as long as external persons do not have access to deleted accounts data, these reports cannot be verified by independent researchers or journalists.

Facebook can also apply a temporary suspension, and in this case the data can often be collected and analyzed. For example, Donald Trump's official Facebook page has been suspended following the Capitol attack on January 6, 2021.[13] Nevertheless the page's data is still present in the CrowdTangle API. Thus, after manually adding this page to the CrowdTangle dashboard, we collected the 6 083 posts it published between January 1, 2020 and June 15, 2021 using the *posts* endpoint.[14] We used the minet command line tool [5] to collect data. We can verify on figure 1 that the *Donald J. Trump* page has not published any content since January 6, 2021, and that this

[10] A list of notable Twitter temporary and permanent suspensions can be found on wikipedia: https://en.wikipedia.org/wiki/Twitter_suspensions.

[11] CrowdTangle is a public insights tool owned and operated by Facebook, that exclusively tracks public content from Facebook public groups and pages.

[12] See the April 2021 report for an example: https://about.fb.com/news/2021/05/april-2021-coordinated-inauthentic-behavior-report/

[13] See https://www.facebook.com/zuck/posts/10112681480907401

[14] (see the endpoint documentation for more details: https://github.com/CrowdTangle/API/wiki/Posts.

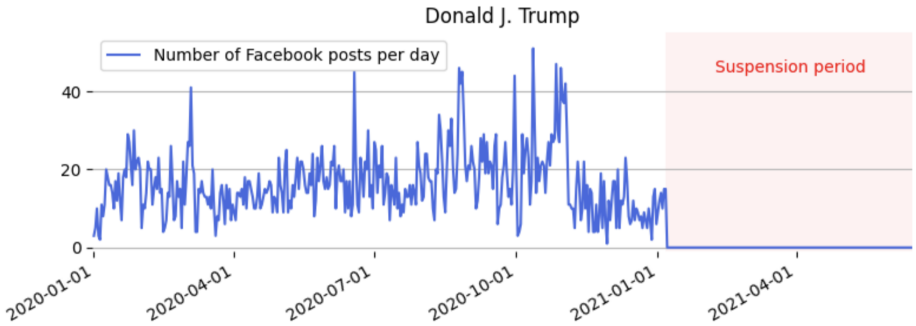

Figure 1: Number of Facebook posts published each day by the Facebook page *Donald J. Trump* between January 1, 2020 and June 15, 2021. The data corresponds to 6 083 posts retrieved from the CrowdTangle API using the *posts* endpoint.

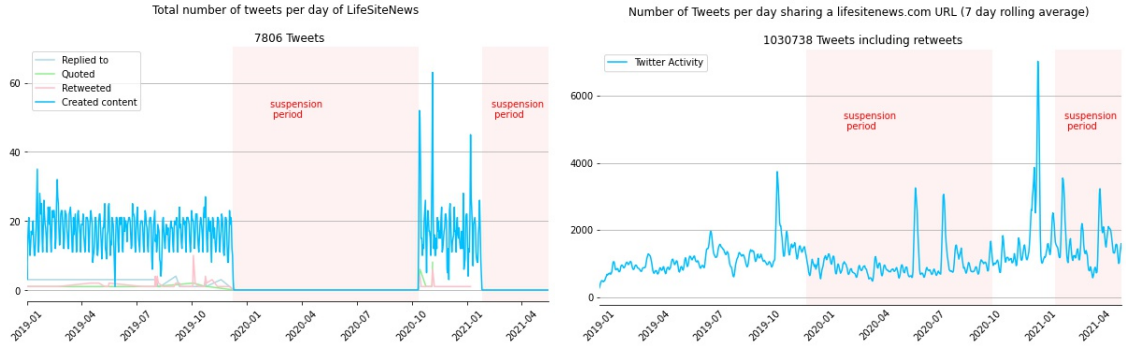

Figure 2: Panel (a): number of Tweets per day of the Twitter account @*Lifesite* linked to the website lifesitenews.com from January, 2019 until April 2021. Panel (b): number of Tweets per day that have shared a lifesitenews.com URL link from January, 2019 until April 2021.

behavior is not consistent with the page's previous activity: an average of 16 posts were published each day on Facebook before the suspension.

2.1.2 Twitter

N.B. Manual decision, not the automatic strike system

In this section, we discuss one case of temporary suspension on Twitter.[15] The Twitter account of the website lifesitenews.com has been suspended for at least two periods of time: from end of 2019 until fall 2020 for 308 days, then again since January 2021 for having violated Twitter Rules[16].

[15] See the official documentation on the Twitter's Help Center regarding account suspension: https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts .

[16] See Lifesitenews's article discussing the reason for the suspension: https://www.lifesitenews.com/news/lifesite-is-dumping-twitter-and-so-should-you. Twitter rules can be found at: https://help.twitter.com/en/rules-and-policies/twitter-rules.

In particular, this website has several failed fact-checks concerning the published articles, according to Iffy.news.[17]. We collected the activity (tweets, replies, quotes, retweets) on their Twitter account $@LifeSite$ via the Twitter API, using the historical search endpoint.[18] We then plotted the number of Tweets, Retweets, Quotes and Replies per day, as shown in panel a of figure 2). The two periods of temporary suspension are clearly observed in the data as the user(s) of the account were not allowed to use the functionalities of the Twitter Platform.

To further assess the impact of this double temporary suspension, we collect via Minet Command Line Tool [5], all the tweets that have shared during the same period a url link containing lifesitenews.com. Panel (b) of figure 2, shows that during both periods of temporary suspension, other users still shared lifesitenews.com links and that the level was only slightly below the tweeting and retweeting levels prior to the first temporary suspension. More specifically, there was an average of 960 tweets (including retweets) per day over the first temporary suspension period of 308 days from December 9, 2019 until October 12, 2020, against an average of 977 tweets (including retweets) per day during the exact same period one year earlier. Finally, panel (b) points towards the limitations of suspending an account to limit the spread of its content.

2.1.3 Youtube

In this section, we turn to the channel's temporary or permanent suspension policy of Youtube. Whenever a channel publishes a video that violates the community guidelines for the first time they will usually receive a warning and the content will be removed. For the second time the channel will start receiving strikes. A first strike results in limiting the access of the Youtube channel for one week, like uploading videos, streaming and other activities. Then a second strike is similar but the suspension will be for two weeks. A third strike results in the termination of the channel. The strike count of a channel lasts 90 days. In the special case, where a video is in extreme violation of the guidelines, the publishing channel may get terminated without a warning.[19] To illustrate the implementation of this policy we provide two examples for the temporary suspension of the following two Youtube channels: One America news Network and Tony Heller.

First, we investigate the temporary suspension case of the Youtube Channel of *One America News channel*. This channel received a first strike on November 24, 2020 for the promotion of a false cure for COVID19.[20] We collected the activity of the channel OANN (video counts, view counts) using the Youtube API v3, between November 2020 and January 2021. For the video counts, we used the playlist endpoint to retrieve the videos uploaded with their publishing date and for the view count we used the IDs of the videos we had from the playlists and via the videos endpoint we retrieved the view counts on June 2021.[21]

In addition, as shown in figure 3 when comparing the week before the suspension from 2020/11/17 to 2020/11/23 and one week after from 2020/12/01 to 2020/12/07 it was found that the view count had decreased by (-45% or -60%), even though the posted number of videos was similar (before 81 videos; after 93 videos or 86 videos). In other words, the suspension period may have a good

[17] See https://mediabiasfactcheck.com/life-site-news/

[18] See the documentation: https://developer.twitter.com/en/docs/twitter-api/tweets/search/api-reference/get-tweets-search-all.

[19] See the "Community Guidelines strike basics". Youtube help, Google Developers, https://support.google.com/youtube/answer/2802032?hl=en. Accessed 21 6 2021.

[20] Reference?

[21] See the Google documentation https://developers.google.com/youtube/v3/docs/videos/list and https://developers.google.com/youtube/v3/docs/playlists/list

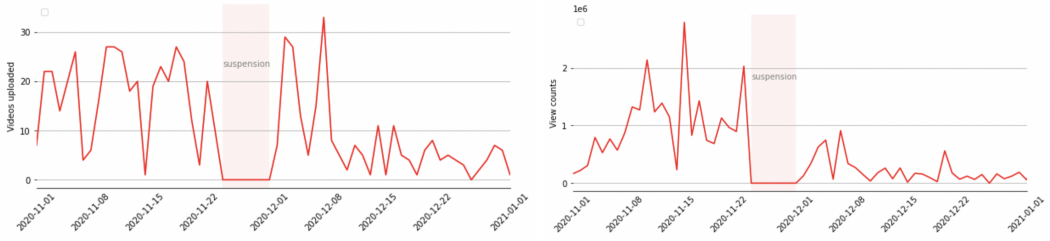

Figure 3: panel (a): Number of Youtube videos uploaded each day by the youtube channel *One America news Network* November 1, 2020 and January 1, 2021. Panel (b): accumulated view counts for videos. The metrics correspond to the videos' publishing date and the data is retrieved from the youtube API with the *playlists* and *videos* endpoints.

Figure 4: Panel (a): Tweet announcing moving to rumble by OANN (Twitter), Twitter ID 1372238828425998336. Panel (b): Tony Heller's tweet after getting suspended from Youtube, Twitter ID 1310703852769796097.

impact on reducing the audience interest or reach to the channel. Besides that, OANN decided to move officially to Rumble on March 17, 2021 as announced on their Twitter account (see figure 4) and their upload activity on their Youtube channel is close to zero since that announcement.

We now turn to our second example, the temporary suspension of the Youtube channel Tony Heller. This channel got its first strike after posting a video about an anti-covid-lockdown doctor getting arrested (see screenshot in figure 4). The suspension period was for one week from September 29 until October 5. We applied the same methods as in the previous example for the data collection. Figure 4 shows the daily number of videos uploaded by the channel. The suspension period can be observed clearly in the historical data of the channel. Moreover, observing the reach of the audience using view counts one week before the suspension starting from 2020/09/22 to 2020/09/28 and one week after the suspension from 2020/10/06 to 2020/10/12 the channel witnessed a drop of -65% and the videos published in the channel were less by -16%.

2.2 Blocking links

A third measure that main stream social media platforms can apply is to prevent users from sharing specific types of content, instead of deleting the content after posting, or users accounts. We will

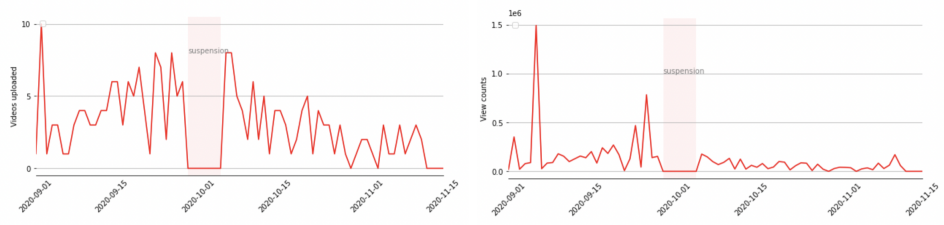

Figure 5: Panel (a): Number of Youtube videos uploaded each day by the Youtube channel *Tony Heller* between September 1, 2020 and November 15, 2020. Panel (b): accumulated view counts for videos uploaded by the same Youtube channel. The date corresponds to the videos' publishing date.

study here how the platforms prevent users from sharing urls coming from specific domain names.

2.2.1 Facebook

The Beauty of life

2.3 Reducing the visibility

Mainstream Social Media platforms can reduce the visibility of the content created or shared by specific users, whenever they violate the platforms' rules. The implementation of this policy varies across platforms and is not easy to verify ex-post. In what follows we provide means to verify this policy on Twitter and Facebook.

2.3.1 Facebook

One of Facebook's measures to regulate misinformation is to reduce the spread of misleading content through ranking. Facebook ranks each post and ad with a relevancy score, a high score leading to a high likelihood of the post or the ad appearing in someone's newsfeed. Facebook can thus make a post or a whole account less visible by decreasing the rank of its content [12]. Such *reduce* action should be visible in the engagement metrics of the posts: if a post reaches less users because of a lower ranking, it is less likely to receive likes, comments and shares.

Owned by Alex Jones, Infowars is usually described as a misinformation website: it is on the Wikimedia global spam blacklist [8], was classified among websites that have posted deceptive content according to FactCheck.org [9], and its factual reporting has been rated *very low* by the Media Bias / Fact Check resource [10]. On May 2, 2019, Facebook announced they would prohibit users from sharing Infowars content unless they are explicitly condemning the material [7].

To verify the measure, we used the CrowdTangle API at the '/posts/search' endpoint (see the documentation [11] for more details) to collect the 37,242 Facebook public posts that shared a link containing 'infowars.com' and that were published between January 1, 2019 and December 31, 2020.[22] The command used can be found here.

[22]We found in the collected data some Facebook posts that did not directly share an Infowars link (but rather a YouTube or Facebook video containing an Infowars link in its description), thus we excluded such posts from our data to keep only the 27,721 posts directly sharing an Infowars link.

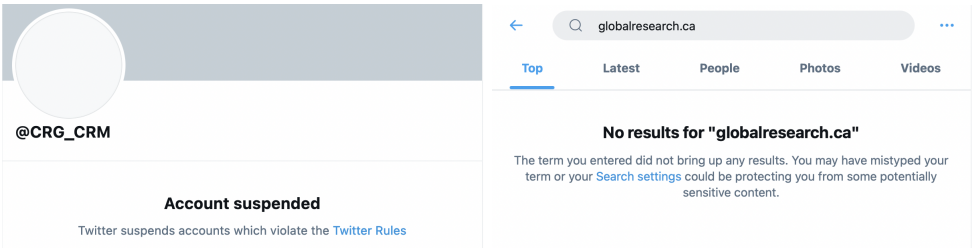

Figure 6: Screenshots taken on June 14, 2021.

The number of public posts sharing an Infowars link remained globally stable in 2019 (Figure X4 top panel), thus Facebook's announced 'ban' did not appear to prevent users from sharing an Infowars link. Interestingly a clear drop in engagement was observed on May 2, 2019 (Figure X4 bottom panel). The number of reactions per post decreased by -94% when we compared the two months before and after May 2, 2019, the shares decreased by -96%, and the comments by -93%.

This would suggest that Facebook's measure taken in May 2019 is actually not a ban, but would rather correspond to a 'reduce' action, as users could still post Infowars links, but these posts generated vastly less engagement. It should be noted that the engagement metrics increased again at the end of 2019 / beginning of 2020, suggesting that the reducing measure may have been lifted a few months after its implementation.

As CrowdTangle is tracking posts only from certain public groups and pages, we also used the '/search/articles' endpoint of the Buzzsumo API, to gather more complete Facebook data. We collected the engagement data for the 14,232 articles crawled by Buzzsumo from the Infowars website between January 1, 2019 and Decembre 31, 2020. The command can be found here. We observe that the articles published after May 2, 2019 received less Facebook engagement than the ones published before (Figure X5), with a percentage change of -97% for the reactions, -59% for the shares and -97% for the comments. An increase in engagement was also observed in 2020. It reinforces the hypothesis that Facebook reduced the reach of posts sharing Infowars links during a few months in 2019.

We found irregularities in the number of Infowars articles collected from Buzzsumo. While Infowars usually publishes 20-30 articles per day, only 53 articles were collected in the 31-day period between June 11 to July 11, 2020. A temporary crawling problem coming from Buzzsumo may have caused this lack of data. Because no database is perfect, we would like to highlight the importance of cross-checking information between different sources when possible.

2.3.2 Twitter

Twitter can take action against a tweet which violates the Twitter rules[23], by limiting its visibility on users' timelines and in search results. To illustrate we provide an example for the website *globalresearch.ca*, which has several failed fact-checks according to iffy.news - a website which provides a database of websites with low factual reporting levels.[24]

The website *globalresearch.ca* is linked to the Twitter account @*CRG_CRM*; that was recently

[23]See the paragraph *Limiting Tweet visibility*: https://help.twitter.com/en/rules-and-policies/enforcement-options.

[24]For *globalresearch.ca* see https://mediabiasfactcheck.com/global-research/.

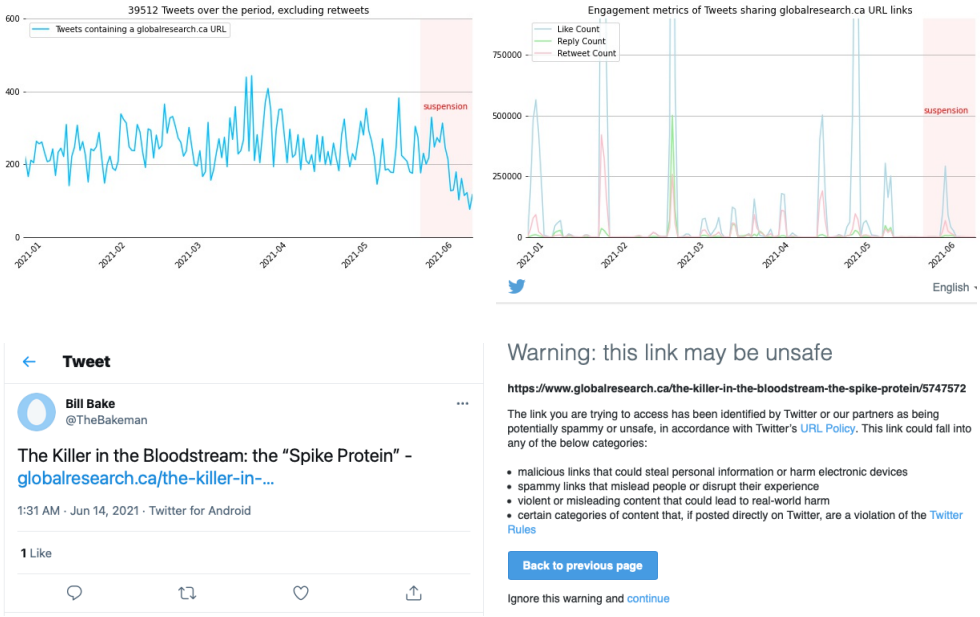

Figure 7

suspended.[25] When a user searches via the twitter search-box a url link of this website, no results appear as shown in the screenshot in panel (*b*) of figure 6, taken on June 14, 2021. To further investigate the possible implementation of a reduced visibility measure, we search via the Twitter API for tweets, excluding retweets, containing the query *globalresearch.ca* from January 1, 2021 until June 10, 2021. As shown in panel (a) in figure 7, we find a strictly positive number of tweets containing the URL link *globalresearch.ca* throughout May 2021 and the first week of June 2021. Hence, the visibility of tweets containing this URL link has been reduced because users can no longer access tweets containing the URL link *globalresearch.ca* via the search box. Nevertheless users are not restrained from posting tweets containing this URL, as shown in the screenshot in panel (*c*) of figure 7, found by taking the tweet ID of one of the collected tweets via the Twitter API. Furthermore, those collected Tweets have strictly positive engagement metrics as shown in panel (*b*) of figure 7. Hence, the users who tweet articles from the *globalresearch.ca* website receive positive engagement from their own followers. Finally, when a user attempts to click on the URL link *globalresearch.ca* contained in the Tweet, a warning message appears and indicates that the link may be unsafe (see screenshot in panel (d) in figure 7).

2.3.3 Youtube

What about something about recommendations ? authoritative content

[25] We noticed the message about the account suspension on May 25, 2021. But to the best of our knowledge, no official communication by Twitter has announced the suspension nor the exact date at which it was implemented. Hence the account may have gotten suspended anytime between April 15, 2021 and May 25, 2021 (see the suspension screenshot in panel (*a*) of figure 6). Furthermore, *globalresearch.ca* has multiple failed fact-checks, see https://mediabiasfactcheck.com/global-research/

2.4 Flags and Notices

2.4.1 Twitter

https://help.twitter.com/en/rules-and-policies/notices-on-twitter

We gathered a set of 3094 URL links of articles which were marked as $False$ by Science Feedback, a not-for-profit organization verifying the credibility of influential claims and media coverage that claims to be scientific. As a second step, we collected (on June 30, 2021) via Minet Command line tool [5] all the tweets that have shared a URL link which belongs to the set of 3094 links marked as $False$. This data collection resulted in 323 938 tweets, excluding retweets. Only 28 tweets were marked with the notice "Get the facts about COVID-19", 5 tweets were marked with "Learn about US 2020 election security efforts" and only 1 tweet was marked with the notice "This claim about election fraud is disputed". Furthermore, we noticed that there exists tweets that share the exact same URL links marked as False, yet only a few get marked by a notice.

Furthermore,

3 Discussion

For a previous research project[26], we searched on CrowdTangle for public accounts sharing specific content associated with misinformation in November 2020, and selected 94 Facebook pages corresponding to our criteria. We then tried to collect these pages' posts in January 2021, and discovered that 11 pages could not be found anymore. This highlights an important issue when studying misinformation trends on Facebook: some data disappears from the CrowdTangle API as accounts are deleted or changed to *private*.

To facilitate the verification in the policy applications, we would generally recommend for the platforms to be more transparent. But too much transparency on how the regulation policies are exactly implemented can actually backfire. For example YouTube is certainly applying an 'information' banner on all videos mentioning Covid and related terms in their title. Misinformation accounts are trying to avoid the official banners by using terms as 'C.O.V.I.D' or 'C O V I D'. If YouTube was totally transparent on that matter and published the list of 'dangerous' words that leads to an information banner, this list would of course help us to understand YouTube's policies but it would also help the misinformation actors to escape the regulation. There is thus a balance between communicating enough so the public can know precisely how the platforms are regulating their content, but without giving too much information that would allow the policies to be bypassed.

There are other ways to collect data from platforms, and besides Buzzsumo, other API are also aggregating data from multiple social platforms. For example Newsguard, blablabla... In this tweet, we can see the interface of XXX being used in this study from a data journalist: https://twitter.com/Shayan86/status/1394742784683298818/photo/2

Recommendation: inform about sources, example inform users that this user shared x failed fact-checks.

https://misinforeview.hks.harvard.edu/article/tackling-misinformation-what-researchers-could-do-with-social-media-data/

[26] reference?

References

[1] David A. Broniatowski, Daniel Kerchner, Fouzia Farooq, Xiaolei Huang, Amelia M. Jamison, Mark Dredze, and Sandra Crouse Quinn. Debunking the misinfodemic: Coronavirus social media contains more, not less, credible content. *mimeo*, 2020.

[2] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on twitter during the 2016 u.s. presidential election. *Science*, 263(274), 2019.

[3] Adam Hughes and Stefan Wojicik. 10 facts about americans and twitter. *Pew Research Center*, 2019.

[4] David Lazer, Matthew Baum, Yochai Benkler, Adam Berinsky, Kelly Greenhill, Filippo Menczer, Miriam Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven Sloman, Cass Sunstein, Emily Thorson, Duncan Watts, and Jonathan Zittrain. The science of fake news. *Science*, 359(6380), 2018.

[5] Guillaume Plique, Pauline Breteau, Jules Farjas, Héloïse Théro, and Jean Descamps. Minet, a webmining cli tool and library for python zenodo. http://doi.org/10.5281/zenodo.4564399. 2019.