Instructions for TTO 2019 Proceedings

Anonymous TTO submission

Abstract

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.

Section 230 in the United States Communications Decency Act[1] 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.[2] 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.[3] 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.[4]

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 spam.[5] 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*.[6] 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, has modified the information

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

[2] See Facebook Transparency Center, Content restrictions based on Local Law: https://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.

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

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

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

[6] 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) [**?**].

ecosystem in terms of production of information and its mediation. In particular, many users can now produce content which contains news related information without having to abide by strict editorial processes that ensure accuracy of information.

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.[7] 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.

In this 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 focus on three platforms: Facebook, Twitter and Youtube. Both Facebook and Youtube are in the top 3 most popular social media platforms in terms of number of users.[8] We further choose Twitter because it is a social networking platform with the most news-focused users, according to the Pew Research center (2019) [**?**].

In this article, 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. Second, we provide simple means to check how those policies are implemented in practice and discuss how to assess their impact, when possible. 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.

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

[8]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/.

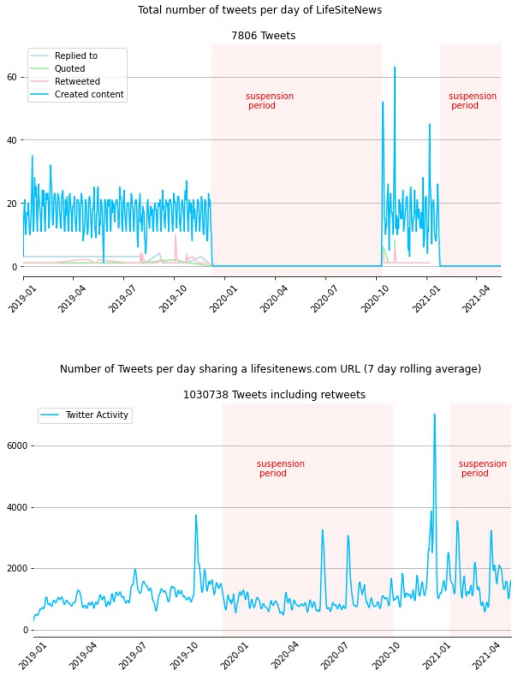

Figure 1: Panel (a): number of Tweets per day of the Twitter account @$Lifesite$ linked to the website lifesitenews.com from January, 2019 until April 2021. Panel (b): number of Tweets per day that have shared a lifesitenews.com URL link from January, 2019 until April 2021.

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.[9] 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 Twitter

In this section, we discuss one case of temporary suspension on Twitter.[10] 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[11]. In particular, this website has several failed fact-checks concerning the published articles, according to Iffy.news.[12]. We collected the activity (tweets, replies, quotes, retweets) on their Twitter account $@LifeSite$ via the Twitter API, using the historical search endpoint.[13] We then plotted the number of Tweets, Retweets, Quotes and Replies per day, as shown in panel a of figure 1). 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[3], all the tweets that have shared during the same period a url link containing lifesitenews.com. Panel (b) of figure 1, 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.

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

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

[11] See Lifesitenews's article discussing the reason of 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.

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

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

2.1.2 Facebook

Unfortunately 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.[14] For a previous research project[15], 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*.

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.[16] 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.[17] 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.[18] We used the minet command line tool [3] to collect data. We can verify on figure 2 that the *Donald J. Trump* page has not published any content since January 6, 2021, and that this behavior is not consistent with the page's previous activity: an average of 16 posts were published each day on Facebook before the suspension.

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

[15] reference?

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

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

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

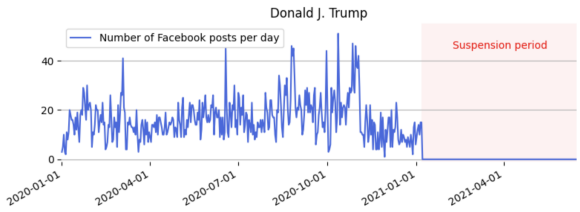

Figure 2: Number of Facebook posts published each day by the Facebook page *Donald J. Trump* between January 1, 2020 and June 15, 2021. The data corresponds to 6 083 posts retrieved from the CrowdTangle API using the *posts* endpoint.

2.1.3 Youtube

In this section, we turn to the channel 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]

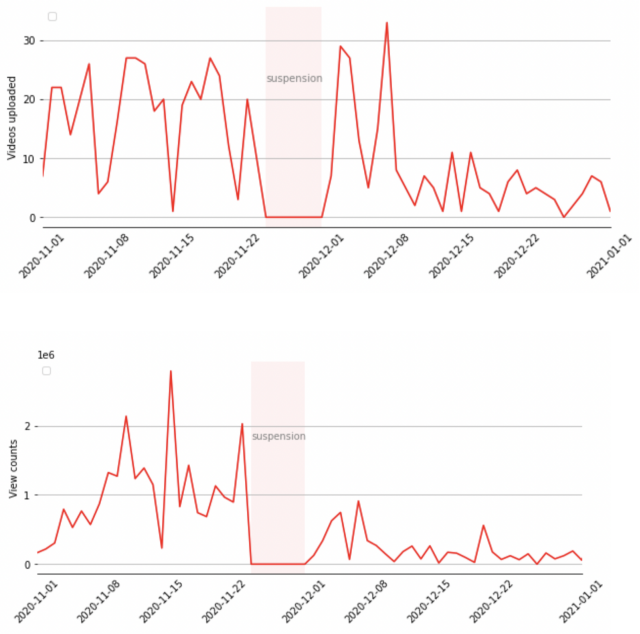

Figure 3: panel (a): Number of Youtube videos uploaded each day by the youtube channel *One America news Network* November 1, 2020 and January 1, 2021. Panel (b): accumulated view counts for videos. The metrics correspond to the videos' publishing date and the data is retrieved from the youtube API with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

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 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 5) 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 5).

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

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

2.2.1 Facebook

The Beauty of life

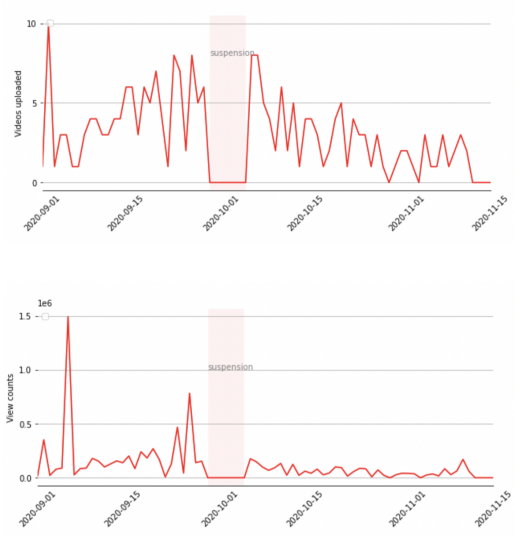

Figure 5: Panel (a): Number of Youtube videos uploaded each day by the Youtube channel *Tony Heller* between September 1, 2020 and November 15, 2020. Panel (b): accumulated view counts for videos uploaded by the same Youtube channel. The date corresponds to the videos' publishing date.

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 accros 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

Infowars

2.3.2 Twitter

Twitter can take action against a tweet which violates the Twitter rules[22], 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.[23]

The website $globalresearch.ca$ is linked to the Twitter account $@CRG\_CRM$; that was recently

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

[23] For $globalresearch.ca$ see https://mediabiasfactcheck.com/global-research/ .

suspended.[24] 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 **??**, 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 6, 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 6, 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 6. 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 6).

2.4 Flags and Notices

3 Discussion

[24] 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 **??**). Furthermore, $globalresearch.ca$ has multiple failed fact-checks, see https://mediabiasfactcheck.com/global-research/

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] 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.

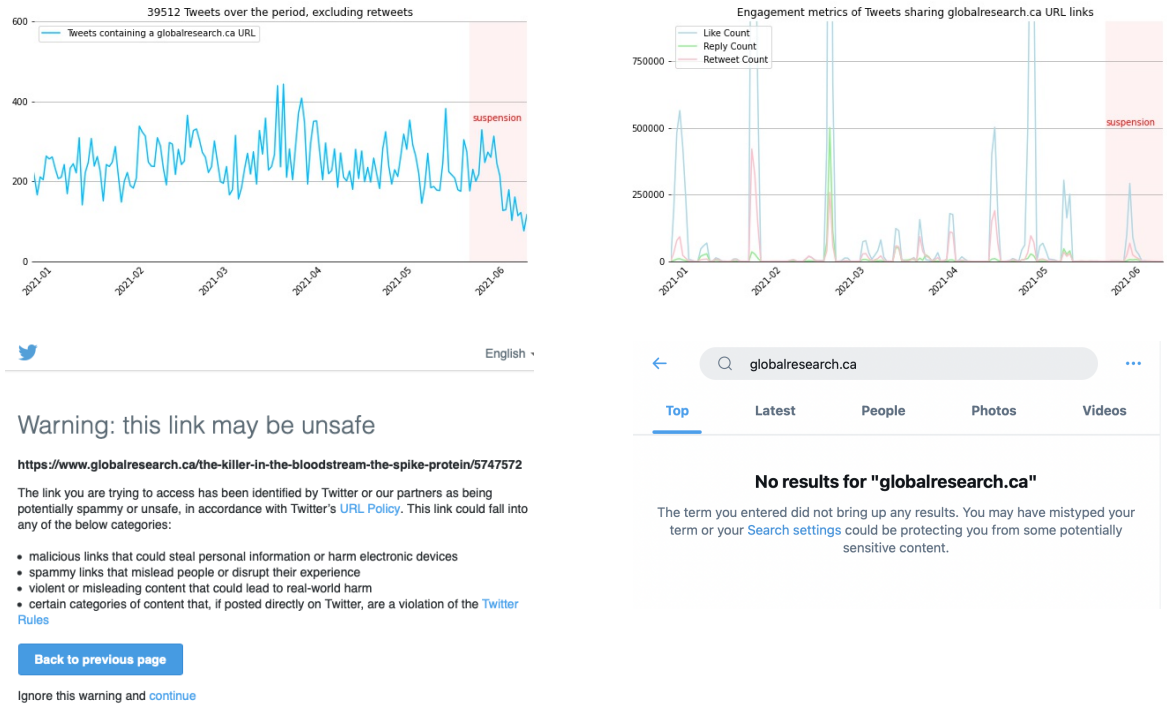
39512 Tweets over the period, excluding retweets
Tweets containing a globalresearch.ca URL
suspension
2021-01
2021-02
2021-03
2021-04
2021-05
2021-06
Engagement metrics of Tweets sharing globalresearch.ca URL links
Like Count
Reply Count
Retweet Count
suspension
English
Warning: this link may be unsafe
https://www.globalresearch.ca/the-killer-in-the-bloodstream-the-spike-protein/5747572
The link you are trying to access has been identified by Twitter or our partners as being potentially spammy or unsafe, in accordance with Twitter's URL Policy. This link could fall into any of the below categories:
malicious links that could steal personal information or harm electronic devices
spammy links that mislead people or disrupt their experience
violent or misleading content that could lead to real-world harm
certain categories of content that, if posted directly on Twitter, are a violation of the Twitter Rules
Back to previous page
Ignore this warning and continue
globalresearch.ca
Top
Latest
People
Photos
Videos
No results for "globalresearch.ca"
The term you entered did not bring up any results. You may have mistyped your term or your Search settings could be protecting you from some potentially sensitive content.

Figure 6