[Policy Brief]

Misinformation on Digital Platforms

March 30, 2022

[**In a nutshell**] We survey the main types of misinformation related interventions by digital platforms and point out what kind of data is available for monitoring and auditing purposes. We further provide a methodology based on currently available data which allows for auditing and monitoring these interventions. We do so by providing a series of examples for different interventions and platforms. Finally, we provide a specific list of recommendations, that could be useful to both digital platforms and regulators, when pushing forward their efforts to tackle disinformation.

Contents

1 What's the issue?

[**Context**] There is a growing concern in society about the spread of disinformation on online platforms and its potential impact on democratic debates, public health and security. To address this concern, online platforms are expanding their rules in order to tackle the spread of misleading or false information. Furthermore, regulators on a national and European level are making progress on the design of legal framework specifically tailored to tackle disinformation.[1] During the COVID-19 global health pandemic, platforms have shown a willingness to ensure the access to reliable health information by implementing further new rules, and have cooperated with regulators on national levels by implementing COVID-19 specific monitoring and reporting programmes. Algorithmic tasks hard to implement for opinions, content etc. but successful for other policy areas.

[**Challenge**] Assessing regularly the impact of misinformation related interventions by online platforms and monitoring their implementation, as well as a careful investigation of the phenomenon of misinformation itself, are necessary safe-guards for democratic societies, with growing digital spheres. Since their early days, online platforms have emerged as digital spaces where information and opinions can circulate freely.[2] The task of ensuring a balance between freedom of expression and access to reliable information regarding the political life or public health, is tremendously intricate. In spite of transparency efforts by digital platforms and new regulatory guidelines, a number of issues remain. Namely, there is a limited access to specific data which would enable the academic community, NGOs, the civil society and data journalists successfully study online misinformation and related interventions.

2 How do platforms action content related to misinformation?

Misinformation related interventions by mainstream platforms are hard to monitor, study or verify by third parties (e.g. academic community, data journalists, NGOs), as in many cases misleading or false content does not qualify as a violation of a given platform's rules and it does not explicitly appear as a separate category in available transparency reports. Algorithmic task hard to implement for opinions, content etc. but successful for other policy areas. False or inaccurate content produced and shared on social networking platforms concerning the political life or public health may have a potentially harmful impact on the society, in the rare event that it goes viral. This gave rise to a set of heterogenous fact-checking policies across mainstream platforms. For example, Facebook has a substantial partnership program with Fact-checking partners certified by the non-partisan International Fact-Checking Network. Facebook uses a number of signals and machine learning models to predict misinformation and surface it to fact-checkers.[3] Twitter seems to have a different approach where they focus on providing context rather than fact-checking[4] and the platform is

[1] See the Chronology of EU's action against disinformation from 2015 to 2021.

[2] The law have encouraged this approach, see: section 230 + e-commerce directive articles 12 and 15

[3] See the section Frequently asked questions: 'How does Facebook use technology to detect potential misinformation?"

[4] To the best of our knowledge, Twitter does not have a page which summarizes its fact-checking strategy. The Twitter Safety Team tweeted on June 3, 2020 the following: "We heard: 1. Twitter shouldn't determine the truthfulness of Tweets 2. Twitter should provide context to help people make up their own minds in cases where the substance of a Tweet is disputed. Hence, our focus is on providing context, not fact-checking." Tweet ID

testing a new system based on the wisdom of the crowds to tackle misinformation (see Twitter Birdwatch). As for YouTube, this platform utilizes the schema.org ClaimReview markup, where fact-checking articles created by eligible publishers can appear on information panels (see Appendix 5 for references).

During the COVID-19 global health pandemic, platforms have upgraded their guidelines to include a set of rules to tackle the propagation of potentially harmful content. Those policies are enforced via existing actions used by the platforms to tackle other rules' violations, such as: labelling content to provide more context or indicate falsehood, publishing a list of terms or topics that will be flagged, suspending accounts, implementing strike systems, reducing the visibility of content, etc. Facebook's strategy to tackle "False News" is three fold : Remove, Reduce, Inform. It is explained via a blog post on the facebook Newsroom. Similarly, Twitter communicates about actions related to misinformation via their Twitter Safety Blog/account. In a post on YouTube Official Blog, the platform explained its "Four Rs of Responsibility" and how it raises authoritative content, reduces borderline content and harmful misinformation. As each platform is a private company, those *new* policies are not coordinated and are implemented in different ways across platforms. However there are common policies against misinformation used by Facebook, Twitter and YouTube, which could be classified into three broad categories: (*i*) temporary or permanent suspension of users, (*ii*) informing users with flags and notices, and (*iii*) reducing the visibility of some content. We compile in table 2, a list of links that redirect to the policies, regulations and transparency centers of Facebook, Twitter and YouTube that we discuss throughout the article.

3 Can we monitor and audit misinformation related interventions with currently available data?

condor

In order to investigate the interventions of mainstream digital platforms regarding misleading or inaccurate content, with currently available data, we take as a starting point a set of URLs that were flagged as False by a fact-checking organization. More precisely, we use the Condor dataset (version June 2021), which contains all the URLs shared at least 100 times on Facebook between January 1, 2017 and February 28, 2021, along with their fact-checking metadata (Messing et al., 2021). We extract all the URLs flagged as False since January 1, 2020 and restrict the set to ten European countries where the links were mostly shared and that appear > 100 times.

illustrations ...

1. pie chart, number of youtube videos within set : available vs deleted
2. Tweets: soft intervention 13 tweets, possibly sensitive 4%, median/average follower count of the users who created the 13 tweets. what else ?
3. Facebook: reduced visibility graph - self-declared groups? but not from the set

Hence, in the present policy brief we explain how to verify mainstream platforms' current actions regarding content with low credibility or false information with data mining. We do so by providing

1267986503721988096.

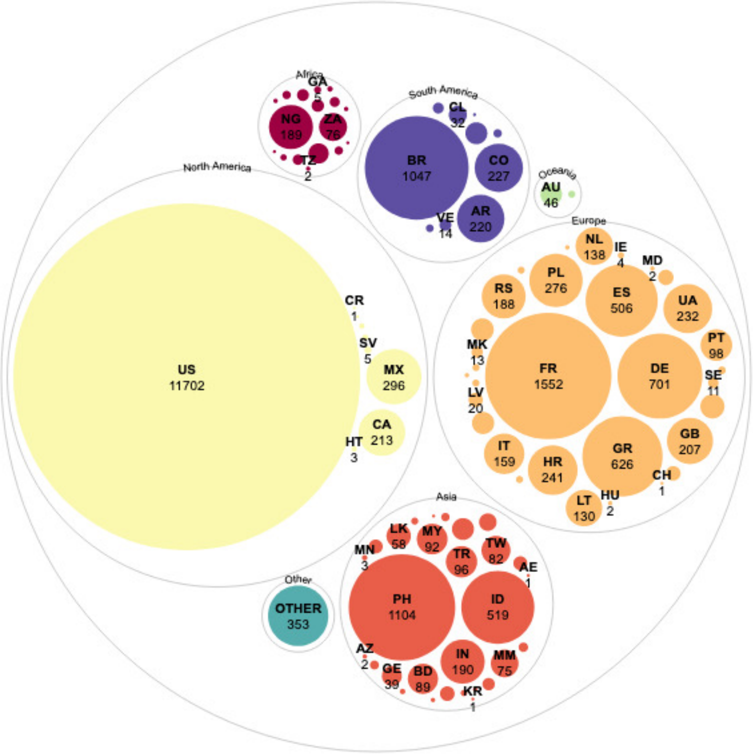

Figure 1: Number of links "fact checked as false" in the Condor Dataset, by country where it was most shared ('public_shares_top_country").

a series of examples for different interventions and platforms. We chose to 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.[5] We further choose Twitter because it is a social networking platform with the most news-focused users, according to the Pew Research center (2019) [1]. To collect data from these three platforms, we either used the APIs (Application Programming Interfaces), or web scraping, i.e. retro-engineering the HMTL code of a web page to extract meaningful data, see Table 1 for a summary. Minet [6], a webmining tool developed by the SciencesPo médialab, was often used, and we scripted our own data mining code when it was necessary.

In this article we have assessed three broad categories of interventions to tackle misinformation by Facebook, Twitter and YouTube: (*i*) suspension of accounts, (*ii*) introduction of flags, labels and information panels and (*iii*) reducing the visibility of content.[6] Furthermore, we have provided a methodology that can be useful for third-party monitoring. The code needed to collect the data and create the figures is available on the following public GitHub repository: https://github.com/medialab/truth-and-trust-online-2021/, and we hope it can be a useful ressource for researchers, NGOs and journalists addressing online misinformation.

	Application Programming Interface (API)	Web Scraping
f	CrowdTangle API and Buzzsumo API	Code created for this article
🐦	Twitter API v2	Minet tw scrape
▶	YouTube API v3	Code created for this article

Table 1: Summary of ressources used for data collection.

Notational variants

We are aware that full transparency regarding operational aspects of misinformation related interventions can backfire. This is because the aimed users might find means to circumvent the interventions. We noticed that some YouTube channels are trying to avoid the likely automated COVID-19 information panels, by using notational variants such as “C.O.V.I.D” or “C O V I D”. Similarly, we came across notational variants on Twitter of the domain name *off-guardian.org* [7] such as *off-guardian. org*, *off-guardian .org*, *off-guardian./org*, *off-guardian*org*.[8]

4 What should policy makers and digital platforms do?

1. For monitoring and auditing tasks, **give access via main APIs, to deleted content due to policy violations** and specify the reason.

[5] 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/.

[6] For a detailed list of interventions of multiple platforms, we invite the reader to navigate through the following airtable compiled by Slatz and Leibowicz (2021) [8].

[7] Here is an example of a failed fact-check of one of the articles of the website *off-guardian.org* : politifact.com/factchecks/2020/jul/07/blog-posting/covid-19-tests-are-not-scientifically-meaningless/

[8] For this case, we failed to understand precisely what kind of intervention Twitter users were circumventing, since they could share in a Tweet an article redirecting to the website *off-guardian.org* with the correct notation and only a warning message would appear when one clicks on it.

Digital platform can remove content that violates their rules or or invite users to delete it themselves via a notification to regain access to the functionalities of their account. In both cases, the data can no longer be accessed via scraping or via the official APIs (Twitter, YouTube, CrowdTangle) for the purposes of scientific research. This hampers the investigation of which content is moderated and why. Moreover data can disappear from platforms over time (see Mazoyer et al. 2020 [2]), hence it is hard to distinguish whether platforms or users are the ones responsible for data removal.

2. For monitoring and auditing tasks, **give access via main APIs to essential metrics, such as:**

   (a) a field indicating the **"reach" of posts on Facebook.**

   The number of people who actually see a given post, cannot be recovered from the CrowdTangle API.[9] Hence, researchers can only build proxies for this metric based on available data.

   (b) a field indicating the **policy violation leading to account suspension** on Twitter, Youtube and Facebook, and the content itself that led to that decision.

   (c) a field indicating **the presence of a banner, an information panel, a notice.**

   To the best of our knowledge, fields related to flags, notices and information panels are not available on the official APIs, with the exception of the *withheld* and *possibly sensitive content* interstitials in the Twitter API v2.

   (d) a field indicating whether **content was signaled and how many times.**

   (e) a field indicating the **number of strikes received by an account** (Twitter, Facebook, Youtube) along with the content that caused the strike.

3. **Allow the access to more data - Facebook Crowdtangle**

   There is a growing restrictions to access APIs (Perriam et al.(2019) [5]). In April 2018, Facebook severely limited access to their official APIs[10] and CrowdTangle is today a reliable source for researchers to access Facebook data. However, it should be noted that CrowdTangle has been recently used by journalists to show that Facebook is spreading biased information[11] and a general suspicion is growing that its access might be shut down in a near future to journalists and researchers.[12]

   Regarding Twitter and YouTube, it is still possible and relatively easy to access their official APIs. Nevertheless the number of Tweets that can be retrieved has decreased between the

[9] See https://help.crowdtangle.com/en/articles/4558716-understanding-and-citing-crowdtangle-data.

[10] See https://about.fb.com/news/2018/04/restricting-data-access/ and http://thepoliticsofsystems.net/2018/08/ /facebooks-app-review-and-how-independent-research-just-got-a-lot-harder/.

[11] See the news article https://www.economist.com/graphic-detail/2020/09/10/facebook-offers-a-distorted-view-of-american-news and the tweet https://twitter.com/facebookstop10. The last link is a Twitter account showing each day the ten most shared link on Facebook, most of them coming from extreme right and misinformation sources. Facebook has reacted to this here: https://about.fb.com/news/2020/11/what-do-people-actually-see-on-facebook-in-the-us/.

[12] See the news articles https://www.nytimes.com/2021/07/14/technology/facebook-data.html and https://www.bloomberg.com/news/articles/2021-08-03/facebook-disables-accounts-tied-to-nyu-research-project.

Twitter API v1 and v2. In the v2, there is a monthly cap of $10,000,000$ Tweets for the academic research track[13], while in the v1 there was no monthly limit and only a limit on the number of requests that could be made in a 15-minute window[14]. We thus noticed that we could get around 10 times more Tweets per month with the v1 when compared to the v2 for one specific collection.

Number of articles on google scholar showing disproportionate access when compared to the size of the platform:

- Twitter (all time) : about 7,830,000 results (0.19 sec)
- Facebook (all time) : about 7,040,000 results (0.06 sec)
- Youtube (all time) : about 6,520,000 results (0.03 sec)
- Telegram (all time, possible mistakes): about 5,970 results (0.02 sec)

4. **Enrich transparency reports with a specific section for misinformation interventions + cross categories**

   When navigating through the categories of policy areas, to the best of our knowledge, misinformation is absent from the data available on the transparency centers of Facebook, Twitter and YouTube. Platforms' official communication about misinformation interventions is scarce and the academic community, NGOs and data journalists, usually discover interventions related to misinformation via monitoring social media accounts related to domain names with several failed fact-checks or via articles in News outlets.

   In certain cases, the platform interventions were not motivated by content moderation to limit misinformation, but rather by other rule violations. For example, the official reason for the suspension of Donald Trump's Twitter account was not misinformation[15] but the violation of Twitter's Glorification of Violence Policy, with his last tweets encouraging the Capitol's riot.[16] Similarly, when Facebook removed four pages related to the InfoWars website, the cited reason was repeated violations of Community Standards: "While much of the discussion around Infowars has been related to false news, which is a serious issue that we are working to address [...], none of the violations that spurred today's removals were related to this".[17] Regarding The Beauty of Life website, its ban was officially due to coordinated inauthentic behavior,[18] which includes using fake accounts that misrepresent one's identity or using methods to artificially boost the popularity of content. According to Facebook, coordinated inauthentic behavior is a distinct phenomenon from disinformation, as "most of the content shared by coordinated manipulation campaigns isn't probably false".[19] Nevertheless, both misinformation and coordinated inauthentic behavior were attested for the Beauty of Life,

[13] https://developer.twitter.com/en/docs/projects/overview

[14] https://developer.twitter.com/en/docs/twitter-api/v1/rate-limits

[15] Although Donald Trump may have spread many misinformation during his governance, see for example: http://allianceforscience.cornell.edu/blog/2020/10/what-drove-the-covid-misinformation-infodemic/.

[16] See https://blog.twitter.com/en_us/topics/company/2020/suspension.

[17] See the article https://about.fb.com/news/2018/08/enforcing-our-community-standards/.

[18] See https://about.fb.com/news/2019/12/removing-coordinated-inauthentic-behavior-from-georgia-vietnam-and-the-us/.

[19] See: https://about.fb.com/news/2019/10/inauthentic-behavior-policy-update/.

and reported to Facebook by the fact-checking organization Snopes.[20] Because misinformation is often associated with creating fake accounts to speed its spread, and incitation to violence, it is sometimes unclear to distinguish between interventions due to misinformation and interventions due to other guideline violations, and more transparent communication on platforms' actions would help monitor the interventions specifically related to misinformation.

Furthermore, it is likely that most interventions used to tackle online misinformation correspond to interventions that were designed for earlier policy areas. To illustrate, YouTube explicitly states [21] that: "Several policies in our Community Guidelines are directly applicable to misinformation", such as the deceptive practices, impersonation and hate speech policies. Hence interventions such as content deletion or account suspension, might be efficient for policy areas for which they were specifically designed, such as the policy area *Adult Nudity & Sexual Activity*. Nevertheless, tackling online misinformation might need adapted interventions. Namely, account suspensions or content deletion motivated by the spread of inaccurate information can be perceived as censorship by the targeted users. In particular, such interventions can generate platform migrations towards new alternative social media platforms (e.g. Telegram, Gab, Minds) with little or no content moderation (see Rogers (2020) [7]).

5. **Specific API for monitoring and auditing and the study of misinformation in general and relevance of policies**

   Moreover, a fact-checking label can lead to unpredicted consequences depending on the perception of the person reading it. On Facebook, Twitter and YouTube, fact-checking messages are currently applied only to some content, while the rest of the content on a given platform remains naturally unlabelled. Such sparse application of warnings can lead to an "implied truth" effect, where users may assume that content without warnings have actually been verified and shown to be true (Pennycook et al. (2020) [4]). Using a different approach, Mosleh et al. (2021) [3] identified Twitter users sharing false news and replied to their false tweets with links redirecting to fact-checking websites. They observed a decrease in the quality and accuracy of the content subsequently shared by the corrected user. These research papers highlight the necessity to investigate potential unintended effects of well-meaning interventions. The methods presented in this paper can help to monitor not only the platforms' actions against misinformation, but also their subsequent consequences in real-life settings.

5 Quick access to community guidelines and platform misinformation policies

[20] See https://www.snopes.com/news/2019/11/12/bl-fake-profiles/, https://www.snopes.com/news/2019/11/12/bl-fake-profiles/ and https://www.snopes.com/news/2019/12/13/facebook-bl-cib/.

[21] See *What policies exist to fight misinformation on Youtube?* youtube.com/intl/en_us/howyoutubeworks/our-commitments/fighting-misinformation/#policies.

Rules	f	facebook.com/communitystandards/recentupdates/
	🐦	help.twitter.com/en/rules-and-policies/twitter-rules
	▶	youtube.com/intl/en_us/howyoutubeworks/policies/community-guidelines/
Rules enforcement	f	transparency.fb.com/data/community-standards-enforcement/
	🐦	transparency.twitter.com/en/reports/rules-enforcement.html
	▶	transparencyreport.google.com/youtube-policy/
Transparency center	f	transparency.fb.com/data/ https://law.yale.edu/yls-today/news/facebook-data-transparency-advisory-group-releases-final-report
	🐦	transparency.twitter.com/en/reports.html
	▶	transparencyreport.google.com/?hl=en
Policy regarding Covid-19	f	https://www.facebook.com/help/230764881494641/
	🐦	help.twitter.com/en/rules-and-policies/medical-misinformation-policy blog.twitter.com/en_us/topics/company/2021/updates-to-our-work-on-covid-19-vaccine-misinformation blog.twitter.com/en_us/topics/company/2020/covid-19
	▶	support.google.com/youtube/answer/9891785
Fact-checking policy	f	facebook.com/journalismproject/programs/third-party-fact-checking/how-it-works
	🐦	x
	▶	support.google.com/youtube/answer/9229632
Fighting misinformation	f	facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news about.fb.com/news/2018/05/hard-questions-false-news/
	🐦	
	▶	youtube.com/intl/en_us/howyoutubeworks/our-commitments/fighting-misinformation/#policies blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/
Strike System	f	See in the following link the section *What is the number of strikes a person or Page has to get to before you ban them?* about.fb.com/news/2018/08/enforcing-our-community-standards/ https://transparency.fb.com/en-gb/enforcement/taking-action/counting-strikes/ (updated July 29,2021)
	🐦	See in the following links the section: *Account locks and permanent suspension* help.twitter.com/en/rules-and-policies/election-integrity-policy help.twitter.com/en/rules-and-policies/medical-misinformation-policy
	▶	https://support.google.com/youtube/answer/2802032?hl=en. Accessed 21 6 2021
Account suspension	f	
	🐦	https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts
	▶	
Flags, Notice and Information Panels	f	https://www.facebook.com/business/help/341102040382165
	🐦	https://help.twitter.com/en/rules-and-policies/notices-on-twitter
	▶	support.google.com/youtube/answer/9004474?hl=en

Table 2: Summary of ressources, last accessed on July 5, 2021.

Further reading

-

References

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

[2] Béatrice Mazoyer, Julia Cagé, Nicolas Hervé, and Céline Hudelot. A french corpus for event detection on twitter. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6220–6227, 2020.

[3] Mohsen Mosleh, Cameron Martel, Dean Eckles, and David Rand. Perverse downstream consequences of debunking: Being corrected by another user for posting false political news increases subsequent sharing of low quality, partisan, and toxic content in a twitter field experiment. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

[4] Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, 66(11):4944–4957, 2020.

[5] Jessamy Perriam, Andreas Birkbak, and Andy Freeman. Digital methods in a post-api environment. *International Journal of Social Research Methodology*, 2019.

[6] 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.

[7] Richard Rogers. Deplatforming: Following extreme internet celebrities to telegram and alternative social media deplatforming: Following extreme internet celebrities to telegram and alternative social media deplatforming: Following extreme internet celebreties to telegram and alternative social media. *European Journal of Communication*, 35(3):213–229, 2020.

[8] Emily Saltz and Claire Leibowicz. Shadow bans, fact-checks, info hubs: The big guide to how platforms are handling misinformation in 2021. *NiemanLab https://www.niemanlab.org/2021/06/shadow-bans-fact-checks-info-hubs-the-big-guide-to-how-platforms-are-handling-misinformation-in-2021/*, June 2021.