Journal of Open Humanities Data

A Dataset of COVID-Related Misinformation Videos and their Spread on Social Media

DATA PAPER

ALEKSI KNUUTILA
ALIAKSANDR HERASIMENKA
HUBERT AU
JONATHAN BRIGHT
PHILIP N. HOWARD

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

This dataset contains metadata about all COVID-related YouTube videos which circulated on public social media, but which YouTube eventually removed because they contained false information. It describes 8,122 videos that were shared between November 2019 and June 2020. The dataset contains unique identifiers for the videos and social media accounts that shared the videos, statistics on social media engagement and metadata such as video titles and view counts where they were recoverable. The dataset has reuse potential for research studying narratives related to the coronavirus, the impact of social media on knowledge about health and the politics of social media platforms.

CORRESPONDING AUTHOR:
Aleksi Knuutila
Oxford Internet Institute, Oxford University, Oxford, United Kingdom
aleksi.knuutila@oii.ox.ac.uk

KEYWORDS:
Coronavirus, misinformation; social media; content moderation; platform policies

TO CITE THIS ARTICLE:
Knuutila, A., Herasimenka, A., Au, H., Bright, J., & Howard, P. N. (2021). A Dataset of COVID-Related Misinformation Videos and their Spread on Social Media. *Journal of Open Humanities Data*, 7: 6, pp. 1–5. DOI: *https://doi.org/10.5334/johd.24*

OVERVIEW

REPOSITORY LOCATION

https://doi.org/10.5281/zenodo.4557828

CONTEXT

Misinformation and conspiratorial claims related to the coronavirus are a problem for stemming the pandemic. Public health depends upon people having accurate knowledge about the severity of the problem, how they can avoid infection and what treatments can help them (Goldacre, 2009). Studies have also shown that believing in conspiracy theories makes people less likely to participate in behaviours that protect their health, such as obtaining vaccinations (Dunn et al., 2017). While all large social media platforms can host misinformation, research suggests that YouTube has played a particularly important role as a source for misinformation related to the coronavirus pandemic (Allington, Duffy, Wessely, Dhavan, & Rubin, 2020).

In April 2020, YouTube's Chief Executive Susan Wojcicki stated that the company was increasing its efforts to remove "medically unsubstantiated" videos, using both automated detection as well as human moderators (Cellan-Jones, 2020). YouTube publishes only aggregated information about the videos that break its Community Guidelines and that are removed. It is, however, possible to gather information about individual removed videos from various public data sources. This dataset describes all videos that circulated on publicly searchable social media and then were removed by YouTube because they contained false information about the coronavirus.

The dataset was created for the Computational Propaganda Project at the Oxford Internet Institute, in order to study the scale of the audience of COVID-related misinformation and its mechanisms of distribution on social media.

METHOD

STEPS

This dataset describes 8,122 YouTube videos that contain COVID-related misinformation. Instead of applying our own inclusion criteria, we identify these videos by following the categorisations made by YouTube itself when removing videos.

We identified COVID-related videos by looking for posts on Facebook, Reddit and Twitter that link to YouTube and that match COVID-related keywords. For Twitter, we used an open access dataset that covered the period from October 2019 to the end of April 2020 (Dimitrov et al., 2020). This dataset was based on a set of 268 COVID-related keywords. We simplified and updated this list of keywords to a total of 71 keywords (Knuutila, Herasimenka, Au, Bright, & Howard, 2020). We used the CrowdTangle service to search for posts on Reddit and Facebook between 1 October 2019 and the 30 June 2020. The dataset will not be updated in the future. CrowdTangle is a database that contains public groups and pages from Facebook and Reddit (CrowdTangle, 2020). It does not contain personal accounts or closed groups.

This search resulted in a list of 1,091,876 distinct videos. We then followed the YouTube link to each video, and where the videos were no longer available we recorded the reason that the YouTube site gave for the video having been removed. With this method, we identified 8,122 COVID-related videos that YouTube had removed because they breached its Community Guidelines.

For these 8,122 videos, we recovered additional information and metadata from other sources, since YouTube itself only published the reason for their removal. Firstly, we recovered the titles and part of the description for all the videos that have been posted to Facebook. The posts on Facebook displayed the original titles and the first 157 characters of the video's description, which we could read by programmatically retrieving every Facebook post. We also queried the Facebook Graph API to get the total number of shares, comments and reactions that the videos had received across the entire platform, including posts to individual profiles and closed groups. Data collection was undertaken in July 2020.

Lastly, we recovered metadata about the videos from the *archive.org's WayBack Machine*, a service that archives the older versions of webpages. Copies of the deleted YouTube pages

were accessible through the WayBack Machine's API in 935 cases. For these videos, we could access the view counts, channel subscriber counts, full descriptions of the videos as well as the video's creation date. In 420 cases, we were also able to approximate how long the video had been visible, by noting the date at which the WayBack Machine had archived the first copy of video's page that stated its removal.

CONTENT OF DATASET

The dataset contains the following information:

- The YouTube links where the videos were viewable prior to their removal.
- The titles, descriptions, and view counts of the videos, where these could be recovered.
- The identification numbers of the YouTube channels where the videos were posted and the channels' subscriber counts.
- A timestamp of when the videos were published and removed, where these could be recovered.
- A link to *archive.org* pages where metadata about the videos and in many cases the videos themselves can be viewed.
- Engagement statistics from Facebook's Graph API for every video, describing overall engagement across the platform.
- ID numbers for Twitter and public Facebook posts linking to the videos.

DATASET DESCRIPTION

OBJECT NAME

covid-misinfo-videos.csv

FORMAT NAMES AND VERSIONS

CSV file

CREATION DATES

Data was collected in July 2020 and covers a period from October 2019 to June 2020

DATASET CREATORS

Aleksi Knuutila, Aliaksandr Herasimenko, Hubert Au, Jonathan Bright, Philip N. Howard

LANGUAGE

English

LICENSE

CC-BY

REPOSITORY NAME

Zenodo

PUBLICATION DATE

30th of November 2020

REUSE POTENTIAL

The dataset is a resource for researchers in the humanities that look to study narratives related to the coronavirus and the communities that produce them. One challenge for such studies is that medical misinformation is ephemeral and often quickly removed from social media platforms.

In many cases, however, it is still possible to view and analyse the videos. *Archive.org* and similar services might hold copies of the videos, and the videos' titles may help find them hosted elsewhere. The project that created the dataset raised questions about how to study the content removal policies of platforms and how to utilize the traces left behind by deleted content, and future work in this area may well suggest productive new methodological approaches.

The dataset can also be reused for research on the extent of misinformation in social media and information diets. One benefit of the dataset is that it is a relatively comprehensive list of COVID-related misinformation videos that were shared publicly in the study period.

FUNDING INFORMATION

The work has been funded by the European Research Council and the Oxford Martin Programme on Misinformation, Science, and Media.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Aleksi Knuutila: conceptualisation, formal analysis, writing.

Aliaksandr Herasimenko: formal analysis.

Hubert Au: formal analysis.

Jonathan Bright: project administration, writing.

Philip N. Howard: project administration, writing.

AUTHOR AFFILIATIONS

Aleksi Knuutila *orcid.org/0000-0002-9874-0079*
Oxford Internet Institute, Oxford University, Oxford, United Kingdom
Aliaksandr Herasimenka *orcid.org/0000-0002-5876-5562*
Oxford Internet Institute, Oxford University, Oxford, United Kingdom
Hubert Au *orcid.org/0000-0002-5655-9773*
Oxford Internet Institute, Oxford University, Oxford, United Kingdom
Jonathan Bright
Oxford Internet Institute, Oxford University, Oxford, United Kingdom
Philip N. Howard *orcid.org/0000-0003-3380-821X*
Oxford Internet Institute, Oxford University, Oxford, United Kingdom

PUBLISHER'S NOTE

This paper underwent peer review using the *Cross-Publisher COVID-19 Rapid Review Initiative*.

REFERENCES

Allington, D., Duffy, B., Wessely, S., Dhavan, N., & **Rubin, J.** (2020). Health-protective behaviour, social media usage and conspiracy belief during the COVID-19 public health emergency. *Psychological Medicine*, 1–7. DOI: *https://doi.org/10.1017/S003329172000224X*

Cellan-Jones, R. (2020). YouTube bans 'medically unsubstantiated' content. *BBC News*. Retrieved from *https://www.bbc.com/news/technology-52388586*

CrowdTangle. (2020). What data is CrowdTangle tracking? Retrieved from *http://help.crowdtangle.com/en/articles/1140930-what-data-is-crowdtangle-tracking*

Dimitrov, D., Baran, E., Fafalios, P., Yu, R., Zhu, X., Zloch, M., & **Dietze, S.** (2020). TweetsCOV19 – A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic. *arXiv:2006.14492 [Cs]*. Retrieved from *http://arxiv.org/abs/2006.14492*. DOI: *https://doi.org/10.1145/3340531.3412765*

Dunn, A. G., Surian, D., Leask, J., Dey, A., Mandl, K. D., & **Coiera, E.** (2017). Mapping information exposure on social media to explain differences in HPV vaccine coverage in the United States. *Vaccine*, *35*(23), 3033–3040. DOI: *https://doi.org/10.1016/j.vaccine.2017.04.060*

Goldacre, B. (2009). Media misinformation and health behaviours. *The Lancet Oncology*, *10*(9), 848. *doi:10.1016/S1470-2045(09)70252-9*

Knuutila, A., Herasimenka, A., Au, H., Bright, J., & **Howard, P.** (2020). *Covid-related misinformation on YouTube: The spread of misinformation videos on social media and the effectiveness of platform policies* (Data Memo). Computational Propaganda Project, Oxford Internet Institute. Retrieved from *https://comprop.oii.ox.ac.uk/research/posts/youtube-platform-policies/*. DOI: *https://doi.org/10.1016/S1470-2045(09)70252-9*

TO CITE THIS ARTICLE:
Knuutila, A., Herasimenka, A., Au, H., Bright, J., & Howard, P. N. (2021). A Dataset of COVID-Related Misinformation Videos and their Spread on Social Media. *Journal of Open Humanities Data*, 7: 6, pp. 1–5. DOI: *https://doi.org/10.5334/johd.24*

Published: 11 June 2021

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.