REVISION: A Cross-Country Comparison of Information and Misinformation Provided by Social Media Influencers on YouTube and Youth's Political Involvement

A Data Management Plan created using DMPonline.be

Creator: Darian Harff

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Grant number / URL: 11F6423N

ID: 200356

Start date: 31-10-2022

End date: 30-10-2024

Project abstract:
Social media influencers (SMIs) are recurrently weaving political issues into their communication, which may be engaging for young followers, but also puts them at risk for misinformation, if SMIs use scientifically unfounded claims. Especially youth, whose political identity is still volatile, may be highly susceptible to SMIs' political (mis-)information. However, existing research has neglected to investigate (1) what kind of political (mis-)information is provided by SMIs (2) how it affects youth's political involvement as well as (3) which skills can help them detect harmful content or avoid persuasion by SMIs. To fill these crucial research gaps, I will use a comparative semi-automated content analysis in two countries to provide, for the first time, a systematic overview of the content of SMI's political information on YouTube, and the degree to which it contains misinformation. Building upon these findings, I will draw from the Differential Susceptibility to Media Effects Model and investigate the underlying mechanisms and boundary conditions of the impact of political (mis-)information by SMIs on youth's political involvement within a longitudinal three-wave panel survey. Findings of this PhD project will have crucial implications for political communication research as well as for political and media education practice.

Last modified: 22-06-2023

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
				Only for digital data	*Only for digital data*	*Only for digital data*	*Only for physical data*
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .cvs,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
survey_BG	Contains survey responses of Flemish youth. **These young people will primarily be asked about influencers they follow who (also) post about politics (e.g., YouTuber 'Acid').** **Based on this data, I identify the most popular political influencers among Flemish youth** and assess predictors of using influencers for political information.	New	Digital	Observational	.sav	<100MB	-
allinfvideos_BG	Contains metadata from all YouTube videos of the influencers **that were nominated/named by Flemish young people in surveys. Includes only the following *publicly available* information: video ID, video title, date of posting, video description, channel ID, channel title, number of views, number of likes, number of shares, number of comments, video link.**	New	Digital	Compiled	.sav	<1GB	-
codebook_BG	Contains coding instructions and variable names.	New	Digital	Other	.pdf	<100MB	-
contentanalysis_BG	Contains codes from coders on categories that align with variables in 'codebook_BG'. Coders code public YouTube videos by influencers with high reach, that they access via the YouTube urls listed in the file 'allinfvideos_BG'.	New	Digital	Other	.sav	<1GB	-
longsurvey	Contains longitudinal survey responses of Flemish or German youth. Based on this data, I assess possible effects of being exposed to political influencers' content and mechanisms behind this relationship.	New	Digital	Observational	.sav	<1GB	-

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

Not applicable.

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

No data collection is planned which is based on experiments.

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes

Self-report data related to adolescents' exposure and relationship to influencers and their political content is collected in this PhD project. Moreover, data about their political views and attitudes (e.g., political engagement) is collected. Data is collected via the survey application Qualtrics. Two applications have already been submitted to the ethical committee and have both received approval from SMEC (projects G-2022-4768-R2(MIN) and G-2022-6123-R4(AMD)). In both of the surveys, participation was anonymous.

Other survey data (for the longitudinal study) may be collected in addition to the data collected in the approved projects and ethical approval will be obtained for this project in the second year of the scholarship. Although this study is longitudinal, we will not have to pseudonymize data, because the survey company we collaborate with will recontact participants for participation in wave 2 and 3 (i.e., we will not assess any identifying information).

The content analysis which is part of this PhD project does not assess any personal data, as the metadata that will be used and which is scraped from YouTube only contains publicly available information (such as video title, number of likes, number of comments), not data of any participants. Moreover, the videos analysed in the content analysis are produced by well-known influencers, who have public accounts on the platform YouTube.
In this PhD project, I also do not, in any way, link survey responses to YouTube videos of those same participants. PLEASE NOTE: In general, no YouTube videos or video metadata of *participants* will be coded or used in any way. Instead, one survey collection is used to identify popular influencers based on participants' responses. The subsequent content analysis is conducted on these *influencer* videos.
In the longitudinal study, I then assess consequences of being exposed to those influencers whose videos were analysed in the content analysis among a sample of adolescents, who were not involved in any of the other studies. The linkage is thus only indirect and involves identifying a connection between characteristics of influencers' content, adolescents' exposure to these influencers and adolescents' political attitudes/behavior (e.g., linking exposure to influencers whose videos have populist characteristics to political cynicism among adolescents). The downloading and handling of the video metadata will also be mentioned in the ethical application for the longitudinal study, which will be obtained in the second year of the scholarship.

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, ...)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

For each study, I will create a separate folder that contains the following information:
1. Folder with all the study documents: Ethical application and approval, examples of informed consent files, as well as the information brochures we gave participants. PDFs of questionnaires will be included.
2. Pre-processing documents: The raw data (anonymized sav files) used within the projects.
3. Documents for data analysis: The processed anonymized data (.R or .sps format). A reproducible syntax/code for the data analysis (either .R or .sps format).
4. Open Science: Documents such as data scripts and datasets with survey responses (but without identifying information) will also be made open access via the Open Science Framework (osf.io), in order to allow reproduction of the data. The folder will never contain sensitive (identifiable) information about participants such as names or email addresses. OSF projects will contain wikis which explain the data, so that reproduction is made possible.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- Yes

We will use metadata from our survey studies including date of participation, duration of survey response, browser type, user language, progress in the survey, question text etc. **In addition, metadata of YouTube videos of the most popular political influencers in Belgium and Germany is downloaded (incl. date of posting, number of likes, comments).**

3. Data storage & back-up during the research project

Where will the data be stored?

All data will be stored on secure KU Leuven servers (I-drive or J-drive). Survey responses will be stored on the encrypted network J-drive, and is only accessible to involved researchers using their employee ID and password. Restricted and sensitive data (i.e., email addresses) will be stored separately on an encrypted password-protected personal I-drive of the KU Leuven for the duration of this project, and can only be accessed by the main researchers of this project. Both drives are backed-up automatically to prevent loss of data.

How will the data be backed up?

The data will be stored on the KU Leuven central servers with automatic daily back-up procedures.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

Yes, as the data files are not large (<1GB), we do not anticipate insufficient storage or backup limitations. In case data storage will exceed 5 GB in WP1, which is not expected, a solution will be worked out in collaboration with KU Leuven ICTS service.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

Both the I-drive and J-drive are fully encrypted and password-protected (only accessible to involved researchers using their employee ID).

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

Currently, there are no costs expected for data storage and backup.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

The data, considering it does not include identifiable information, will be retained for at least 10 years, in line with the KU Leuven RDM policy.

Where will these data be archived (stored and curated for the long-term)?

The data will be stored on the university's central servers (with automatic back-up procedures) for at least 10 years, in line with the KU Leuven RDM policy.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

All datasets will be archived on the servers of KU Leuven. Possible expenses to host these servers are always covered by the research group, but are expected to fall within normal cost ranges (based on the volume of the datafiles).

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository

Datasets and scripts will be made available. The codebook of my content analysis will also be made available.

If access is restricted, please specify who will be able to access the data and under what conditions.

Access is not restricted.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- No

Where will the data be made available? If already known, please provide a repository per dataset or data type.

Data (trimmed datasets and data scripts) will be made available on the Open Science Framework under a CC-BY license alongside full data documentation in the wiki.

When will the data be made available?

- Upon publication of the research results.
- Previous to the publication for peer-review via a view-only link that is only shared with reviewers.

Which data usage licenses are you going to provide? If none, please explain why.

We will make our data available under a CC-BY license.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

DOIs may be created which lead to the OSF pages of the respective project.

What are the expected costs for data sharing? How will these costs be covered?

None.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Darian Harff

Who will manage data storage and backup during the research project?

Darian Harff

Who will manage data preservation and sharing?

Darian Harff

Who will update and implement this DMP?

Darian Harff