Interventions against polarisation in society for TRUSTworthy social media (iTRUST)

A Data Management Plan created using DMPonline.be

Creators: Liesbeth Allein, n.n. n.n., Marie-Francine Moens

Affiliation: KU Leuven (KUL)

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

Template: FWO DMP (Flemish Standard DMP)

Principal Investigator: Liesbeth Allein, n.n. n.n. , Marie-Francine Moens

Data Manager: Liesbeth Allein, n.n. n.n., Marie-Francine Moens

Project Administrator: n.n. n.n., Marie-Francine Moens

Grant number / URL: G0L0822N

ID: 199743

Start date: 01-11-2022

End date: 31-10-2025

Project abstract:

Digitalisation is rapidly transforming our societies, transforming the dynamics of our interactions, transforming the culture of our debates. Trust plays a critical role in establishing intellectual humility and interpersonal civility in argumentation and discourse: without it, credibility is doomed, reputation is endangered, cooperation is compromised. The major threats associated with digitalisation - **hate speech and fake news -** are violations of the basic condition for trusting and being trustworthy which are key for constructive, reasonable and responsible communication as well as for the collaborative and ethical organisation of societies. These behaviours eventually lead to **polarisation**, when users repeatedly attack each other in highly emotional terms, focusing on what divides people, not what unites them.

Focusing on two timely domains of interest – gender equality and public health – iTRUST will deliver (i) the largest ever dataset of online text, annotated with features relevant for **ethos, pathos and reframing**; (ii) a new methodology of large-scale comparative trust analytics to detect **implicit patterns and trends** in hate speech and fake news; (iii) a novel **empirical account** of how these patterns affect polarisation in online communication and in society at large; and (iv) AI-based applications that will transfer these insights into **interventions** against hate speech, fake news and polarisation. Given the relevance for the knowledge-based society, the project puts great emphasis on outreach activities and users' awareness in collaboration with media, museums and other partners.

Last modified: 17-05-2023

Interventions against polarisation in society for TRUSTworthy social media (iTRUST)
FWO DMP (Flemish Standard DMP)

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.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		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	
IM	Corpus of 2,000 sentences annotated with moral judgments and reader interpretations	Reuse existing data	Digital	Compiled/aggregated data	.csv	<100MB	
Polaris	Corpus of tweets (120K words) annotated for ethos, pathos, logos	Reuse existing data	Digital	Compiled/aggregated data	.csv	<1GB	
Framing	Corpus of sentences annotated for frames	Reuse existing data	Digital	Compiled/aggregated data	.csv	<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:

Existing data have not been formally published yet, so identifier is currently unavailable.

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

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.

- No

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).

Developed software will contain readme files and documentation. The default is to make the software publicly available on the software page of our lab (https://liir.cs.kuleuven.be/software.php) and duplicated on the software page of the project. We will archive code according to the GitHub standards, which is commonly used in the community. When code is openly released, the corresponding link to the software page is provided in the publication. This way, the public code and data is findable through the corresponding publication, the LIIR software page, and the project software page. Exchanged code must be accompanied by a README file to explain how to use it.

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.

- No

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

Where will the data be stored?

In our LIIR lab data is stored across the following categories of directories:

- Individual Working Directories of researchers on LIIR servers or local machines (WD): e.g., server arwen. The servers are installed with a RAID setup.
- Dedicated Backup Directories on the LIIR servers (BD).
- Separate external Hard Disks (HD).
- The LIIR SoftwareWebpage (https://liir.cs.kuleuven.be/software.php) and the Project's Software Webpage.

How will the data be backed up?

We have dedicated back-up directories on the servers of the LIIR lab and back-ups will also be stored on the university's 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, we have dedicated backup directories on our servers and backup capacity can be extended on the university's central servers.

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

1 Storage Directories
In our LIIR lab data is stored across the following categories of directories:

- Individual Working Directories of researchers on LIIR servers or local machines (WD): server arwen.
- Dedicated Backup Directories on the LIIR servers (BD).
- Separate external Hard Disks (HD).

2 Access Rights to the Data
There are five levels of access rights:
1. Prof. Moens: has access to all storage directories: WD, BD, HD.
2. Systemgroup Staff (including LIIR representative): has access to all WD, the BD.
3. Individual LIIR researchers: have access only to the data they collected or created themselves within the WD, and the BD.
4. Publicly available on request.
5. Publicly available for download on siftware webpage or project's webpage.

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

Data storage is done on our own servers, so there is no extra cost. For backups we will use the servers of KU Leuven (https://icts.kuleuven.be/sc/english/storage/largevolume storage) with a cost of 104.42€/TB/year.

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...).

After publication of each research paper, the 1st author is responsible for saving a complete, documented backup of the code and data needed to reproduce the results reported in the paper. This documented backup is stored at the BD or on HD, depending on privacy sensitivity, and file size, and will be kept for a period of at least 5 years after the project contract terminates. This way, when a researcher departs, a documented backup of the code and data of all published research is stored, ensuring long term reproducibility. This will be checked before the researcher will obtain the PhD degree or before a postdoctoral researcher leaves the department.

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

A complete backup when smaller or equal than 100 GB is archived under: arwen.cs.kuleuven.be:/export/home2/NoCsBack/hci/data management/phds/
Otherwise, a partial backup is stored in the dedicated phd directory, and a complete backup on an external hard disk.

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

The database of sofware that will be compiled will be hosted on the servers of the lab at no extra costs. For the costs of backups see above.

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

The software page is hosted on a departmental server and can be accessed publicly at: http://liir.cs.kuleuven.be/software.php
and will also be made available via the website of the project.

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

Not Applicable

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.

The data will be made available at the software page of the LIIR group: http://liir.cs.kuleuven.be/software.php
and will also be made available via the website of the project

When will the data be made available?

Upon publication of the research results.

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

This software will be released under an MIT license. For a commercial license we request to contact the responsible of the LIIR lab.

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

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

We do not expect extra costs.

6. Responsibilities

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

Marie-Francine Moens, Victor Milewski and Liesbeth Allein

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

Victor Milewski and Liesbeth Allein

Who will manage data preservation and sharing?

Marie-Francine Moens and Victor Milewski

Who will update and implement this DMP?

Marie-Francine Moens bears the end responsibility of updating & implementing this DMP

Interventions against polarisation in society for TRUSTworthy social media (iTRUST)
Application DMP

Questionnaire

Describe the datatypes (surveys, sequences, manuscripts, objects …) the research will collect and/or generate and /or (re)use. (use up to 700 characters)

- Reuse existing data

We plan to use the existing annotated datasets in our research. The IM dataset (.csv file) consists of 2,000 sentences from Reddit posts. The sentences are annotated with generated interpretations of the sentences and labels for reader attitude and perceived moral judgments. Moral judgment characteristics consist of character trait descriptions in natural language, social evaluation label (positive/negative) and character trait classification in Virtue Ethics, a well-adopted moral theory framework. We will also work on datasets that are developed by two other partners in the iTRUST consortium. The POLARIS dataset (.csv file), generated by Warsaw University of Technology, consists of tweets on vaccine hesitancy and climate change (120,000 words) annotated with philosophical concepts related to ethos, pathos and logos. The dataset (.csv file) produced by the Università della Svizzera-Italiana, segments sentences in frames, where entity relations will be made explicit.

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)

Marie-Francine Moens and Liesbeth Allein will be responsible for data documentation and metadata. Victor Milewski and Liesbeth Allein will be responsible for data storage and back up during the project. Marie-Francine Moens and Liesbeth Allein will be responsible for ensuring data preservation and reuse.
Sufficient storage capacity is guaranteed during and after the research project and data will be stored on our own servers.

What’s the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)

NA

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)

NA

Which other issues related to the data management are relevant to mention? (use up to 700 characters)

None

Interventions against polarisation in society for TRUSTworthy social media (iTRUST)
DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- Not applicable

Interventions against polarisation in society for TRUSTworthy social media (iTRUST)
GDPR

GDPR

Have you registered personal data processing activities for this project?

- Not applicable