

CON-NET: New Collected Data - CHIST-ERA format

Version 0

Description

Online misbehaviour means that the information received via social media cannot be taken at face value. Sharing of misleading, out of context or simply false information, and coordination of behaviours ("brigading") distorts the quality of information we receive online, regardless of whether the misbehaviour is intentional. Additionally, by design, each user of modern internet products (e.g., search; social media) is exposed to a unique view of the information landscape without necessarily realizing that this is the case. Every user should be able to answer the questions: "How am I being misinformed?" and "How does my unique position shape the information I receive?" Leveraging a multilayer network approach to detect misbehaviour and describe users' positions in the dynamic information space, the CON-NET project will enable users to answer these questions for themselves or indeed any other network entity.

There are many different factors that make detection and mitigation of online misbehaviour a challenge. The complexity of the networks themselves, the role of automated agents (bots), and coordination to create misleading signals, the motivations behind misbehaviour and the types of misbehaviour all play significant roles. The medium through which misinformation can spread is also an important feature, adding further layers to the complexity. Doctored images and videos can spread through social media along with the ideas and memes of misinformation.

CON-NET will apply machine learning techniques to process the vast amounts of data and identify trends, signals, and suspicious behaviours as well as misbehaving entities, and we will further augment this with a visual analytics approach, including a human in the loop to provide context and human understanding of the concept of misinformation. We will address the complexity of online social media networks by taking a multilayer network approach to their analysis. The aspects for the network will characterise the different platforms, types of entities and interactions,

language, media, as well as how they evolve over time. The multilayer network will be annotated via the machine learning and visualised via an online platform. This platform will push back the frontiers of multilayer network visualization to allow an end user to better understand the source and impact of online misbehaviour and misinformation.

CON-NET requires a large amount of online social media data, from multiple platforms, with each of the partners bringing different expertise and analysis goal. The data will need to be stored processed analysed and visualized. This data management plan will ensure that all data will be stored and managed appropriately

Funder

CHIST-ERA||CHIST-ERA

Grant

CHIST-ERA: Foundations
for Misbehaviour
Detection and Mitigation
Strategies in Online Social
Networks and Media
(OSNEM)

Researchers

Organizations

Sabanci University, Vrije Universiteit Brussel (VUB),
Luxembourg Institute of Science and Technology,
Aalto University

Datasets

Title: [CON-NET: New Collected Data - CHIST-ERA format](#)

Template: [Science Europe](#)

Online misbehaviour means that the information received via social media cannot be taken at face value. Sharing of misleading, out of context or simply false information, and coordination of behaviours ("brigading") distorts the quality of information we receive online, regardless of whether the misbehaviour is intentional. Additionally, by design, each user of modern internet products (e.g., search; social media) is exposed to a unique view of the information landscape without necessarily realizing that this is the case. Every user should be able to answer the questions: "How am I being misinformed?" and "How does my unique position shape the information I receive?" Leveraging a multilayer network approach to detect misbehaviour and describe users' positions in the dynamic information space, the CON-NET project will enable users to answer these questions for themselves or indeed any other network entity.

There are many different factors that make detection and mitigation of online misbehaviour a challenge. The complexity of the networks themselves, the role of automated agents (bots), and coordination to create misleading signals, the motivations behind misbehaviour and the types of misbehaviour all play significant roles. The medium through which misinformation can spread is also an important feature, adding further layers to the complexity. Doctored images and videos can spread through social media along with the ideas and memes of misinformation.

CON-NET will apply machine learning techniques to process the vast amounts of data and identify trends, signals, and suspicious behaviours as well as misbehaving entities, and we will further augment this with a visual analytics approach, including a human in the loop to provide context and human understanding of the concept of misinformation. We will address the complexity of online social media networks by taking a multilayer network approach to their analysis. The aspects for the network will characterise the different platforms, types of entities and interactions, language, media, as well as how they evolve over time. The multilayer network will be annotated via the machine learning and visualised via an online platform. This platform will push back the frontiers of multilayer network visualization to allow an end user to better understand the source and impact of online misbehaviour and misinformation.

CON-NET requires a large amount of online social media data, from multiple platforms, with each of the partners bringing different expertise and analysis goal. The data will need to be stored processed analysed and visualized. This data management plan will ensure that all data will be stored and managed appropriately

Dataset Description

1.2.2 How will new data be collected or produced and/or how will existing data be re-used?

1.2.2.2 Explain which methodologies or software will be used if new data are collected or produced

In the project, we will use social media posts collected from APIs or publicly accessible repositories, media content such as images and video. To begin with the target platform are twitter and reddit. The data can be divided into two broad categories. (1) Private data which is owned by third parties and accessed under strict license conditions or is otherwise sensitive (e.g. social media data) and (2) open data which can be shared in the consortium and to the public (e.g. data from Wikipedia). This category will also include data produced as an outcome of the research process which is aggregated or processed in a way that it is no longer a privacy concerns or license conditions. For example, summary statistics of various of private data sets can belong to this category. Determining to which category the data belongs to is done by following the principle of “as open as possible, as closed as necessary”. All open data will be aligned with the FAIR principles [Findability, Accessibility, Interoperability and Reusability; (Wilkinson, Dumontier and Aalbersberg, 2016)] as detailed below. Raw master data will be archived in its original format. Various data formats needed for research, collaboration, and dissemination are derived from master data, and this is done in simple open formats (e.g. csv, xml) to ensure that data is interoperable with all standard software packages, and reusable even in the distant future. All raw data will be archived in its original format (e.g., the format at which the tweets are provided to us), except in cases with it is necessary to pseudo anonymised the data immediately for GDPR purposes. Data will processed (e.g.,pseudo anonymised (if required), removing unrequired values to perform data minimization) and cleaned to master data in plain text formats such as comma separated values (csv). The various data formats and versions needed for research, collaboration and publicly sharing the data will be produced from the master data. In the published data we will use simple

open formats (such as csv or json) to ensure easy convertibility for use with all statistical and network analytic software packages, and to make sure that the data is usable even in the distant future. Based on similar projects in the past, the total space usage is expected to be less than 10 TB including all derived data. Overall, in data handling and managing procedures, we will follow the guidance for the research data management provided by the consortium institutions.

Custom methodology described above

1.2.2.3 Are there any constraints on the re-use of existing data?

No

1.2.2.4 Explain how data provenance will be documented

Other

We plan to use a data repository with version control features (e.g. GIT) and each data set will contain data describing its provenance (see initial data description above)

1.2.3 What data (for example the kind, formats, and volumes), will be collected or produced?

1.2.3.2 Give details on the data types

- textual (documents)
- image
- audio
- video

Comment:

We will download accounts and online posting data from online social media sites through the service providers APIs. Online posting Data will be pseudo anonymized immediately before being stored. The image videos and audio data will be processed using machine learning to characterize it in terms of misinformation.

We are not storing unprocessed raw online social media textual data as social media data if doing so will violate GDPR. as precautions need to be taken as it contains personal data. However we do consider our Pseudo-anonymized & minimised data sets to be primary data (which we refer to master data above) and not secondary data.

1.2.3.3 Give details on the data format

Other

Comment:

We will store the online social media in json formatted text files and in data bases for processing

2.3.2 What metadata and documentation will accompany the data?

2.3.2.2 Indicate which metadata will be provided to help others identify and discover the data

We will provide English language description of the fields stored in textual data

Comment:

Our primary data will contain meta data to support processing of data in the project. It will be encrypted and stored securely. We will not be providing others the ability to identify and discover this data as it contains personal data in the message text. To do so would be considered a massive violation of GDPR (especial given to number of online social media accounts). Where possible for publications and scientific outputs we will provide anonymized version of the data if appropriate, however we will not put any of the project partners at risks of being responsible for violations of European law

2.3.2.3 Indicate which metadata standards will be used

We will use a simple human readable format.

Comment:

The FAIR data principles will guide our data documentation. The filenames will contain appropriate descriptive information (e.g., an abbreviation for the method and data set) so that it will be easy to recognise them afterwards. The folder structure will separate the different data sets. Data sets will be documented with descriptive metadata (e.g., title, year of publication, creator, description, keywords, etc.), which ensures the understandability and findability of the data in the future. If needed, README files will be created for data sets to ascertain their reusability, reading and interpretation. When applicable, Git is used for software version control, documentation, and preservation.

2.3.2.4 Indicate how the data will be organised during the project

We will use version control software (e.g. GIT). Data from different sources will be kept in different repositories

2.3.2.6 Consider how this information will be captured and where it will be recorded

'readme' text file

Comment:

Each GIT project will have a readme.md file in its root folder describing the data set

2.3.3 What data quality control measures will be used?

2.3.3.1 Explain how the consistency and quality of data collection will be controlled and documented

representation with controlled vocabularies

Comment:

We are collecting massive amount of online social media data. The datasets have undergone or will undergo quality control by the party that generated it. In processing the data and validation of the results, standardized protocols are used, and periodically reviewed by the PIs. The journals publishing data will perform their own quality checks as part of the peer review process. The personnel in the project will participate in training regarding data when appropriate.

3.4.2 How will data and metadata be stored and backed up during the research?

3.4.2.2 Describe where the data will be stored and backed up during research activities and how often the backup will be performed

per Week

The data will be stored on secure servers at partner institutions. LIST will be the primary partner for data storage

Comment:

LIST server will be backed up on a weekly basis. When partners store data, they will implement their own regular backup policies.

3.4.3 How will data security and protection of sensitive data be taken care of during the research?

3.4.3.2 Explain how the data will be recovered in the event of an incident

Loaded from backup

3.4.3.3 Explain who will have access to the data during the research and how access to data is controlled, especially in collaborative partnerships

Only CON-NET consortium members, with restrictions on the partners based outside of Europe

Comment:

Data is stored in group-specific directories with per-person access control. Access rights via groups is managed by the institutions' IT services, but data access is only provided upon request of the PI. All data is made available and transferred only through secure, encrypted, and password-protected systems: it is impossible for any person to get data access without a currently active user account, password, and group access rights. Backups are also kept confidential. All data is securely deleted at the end of the life of the data that is determined for each data set separately. Copies of any sensitive data are not made to unsecure systems (such as non-encrypted laptops). During data analysis, the data will be accessible only to the researchers involved in the research project. Any work done with collaborators accessing non-public data will be done in the consortium institution servers, who will need to get access to the institution's data infrastructure following the guidelines set out by the institution. This ensures that no data will be copied outside of the secure computing platforms.

3.4.3.4 Describe the main risks and how these will be managed

personal data

No personal data will be shared with anybody outside of the project and will be stored encrypted at rest

Comment:

GDPR. Especial care will be given to the GDPR regulation on the transfer of personal data outside of EU as the consortium also includes a member outside of the EU. (notably, Turkey). We will follow data management strategies and rules of set out by the national funding agencies of the consortium partners. The consortium institution's data management guidelines and research ethics committees are consulted when appropriate. When appropriate, we would like to distribute as much anonymised data as possible for reuse under a Creative Commons BY or similar license. When publication of the data in full is not possible, we aim to publish the exact procedures used for processing the data. We will use the MIT license or a similar open license when publishing program code.

3.4.3.5 Explain which institutional data protection policies are in place

Security Measures and Policies implemented by LIST (institution of project Coordinator): 1 - Information security policies 2 - Organization of information security 3 - Human resources security 4 - Asset Management 5- ACCESS CONTROL 6 - Cryptography 7 - Physical and environmental

security 8 - Operations security 9 - Communications security 10 - System acquisition, development, and maintenance 11 - Supplier Relationships 12 - Information security incident management 13 - Information security aspects of business continuity management 14 - Compliance

4.5.2 Personal data

4.5.2.2 If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

All parties of the project consortium are signing a data sharing agreement

4.5.2.3 Explain whether there is a managed access procedure in place for authorised users of personal data

yes

4.5.2.2 Data ownership and accessibility

4.5.2.2.2 Who will be the owner of the data

Whoever collects and creates a data set will be its owner

4.5.2.2.3 Explain what access conditions will apply to the data?

- ethical
- legal

4.5.2.2.4 Will the data be openly accessible, or will there be access restrictions?

Access restrictions

4.5.2.2.5 What kind of access restrictions?

password

Comment:

Data is stored in group-specific directories with per-person access control. Access rights via groups is managed by the institutions' IT services, but data access is only provided upon request of the PI. All data is made available and transferred only through secure, encrypted, and password-protected systems: it is impossible for any person to get data access without a currently active user account, password, and group access rights. Backups are also kept confidential. All data is securely deleted at the end of the life of the data that is determined for each data set separately. Copies of any sensitive data are not made to unsecure systems (such as non-encrypted laptops). During data analysis, the data

will be accessible only to the researchers involved in the research project. Any work done with collaborators accessing non-public data will be done in the consortium institution servers, who will need to get access to the institution's data infrastructure following the guidelines set out by the institution. This ensures that no data will be copied outside of the secure computing platforms.

4.5.2.3 Intellectual property rights

4.5.2.3.2 Are intellectual property rights affected?

Yes

4.5.2.3.3 Explain which intellectual property and how will they be dealt with

Intellectual property rights are defined in the consortium agreement

As described in the consortium agreement

4.5.2.4 Third-party data restrictions

4.5.2.4.2 Indicate whether there are any restrictions on the re-use of third-party data.

Online social media platform APIs each provided their own individual terms and conditions

4.5.4 Ethical issues

4.5.4.2 What ethical issues and codes of conduct (CoC) are there, and how will they be taken into account?

Privacy, persecution of individuals, persecution of researchers

- Anonymisation
- Privacy constraints and applicable ethical norms
- Informed consent statements
- Privacy policies

- National laws

Comment:

The project will provide an ethics policy document providing full details.

5.6.2 How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

5.6.2.2 Explain how the data will be discoverable and shared

other

Comment:

When appropriate, we would like to is to make as much of the data available given the copyright and privacy restrictions (see section 2). Appropriate data, together with metadata, will be published as much as possible with the relevant articles or at the same time as the articles are published but in separate repositories. In addition, metadata and relevant links to the repositories will be published in services indexing data. The deposited data will be supplied with the required standard metadata to ensure reusability and ensuring reproducibility of results in publications. Possible data related to the published materials can be made available in Zenodo or similar repositories (e.g., the Open Science Framework) under the Creative Commons license CC BY or similar license. The repositories also provide persistent identifiers (e.g., DOI, URN) to promote citations. The most suitable archives will be selected out of these for each data set. We will use suitable paper pre-print services, e.g., the ArXiv.org e-Print archive, whenever needed.

5.6.2.3 Outline the plan for data preservation and give information on how long the data will be retained

2030-12-31

Comment:

The private data will be stored in within the consortium institutions' infrastructures described in section 4, and it will be stored until the end of the project with possibility of extending the lifetime of the data. The public research data will be archived long term in the Zenodo data repository or

similar repositories. The Zenodo service provides long-term and easily accessible storage with DOI identifier. Zenodo stores data in CERN Data Center. Both data files and metadata are kept in multiple online and independent replicas. Zenodo is recommended by FAIRsharing.org, and follows FAIR principles, see their Principles page.

5.6.2.4 Explain when the data will be made available

upon publication

Comment:

Data will only be made available if it can be done so under GDPR, and in an ethically valid manner

5.6.2.6 Will exclusive use of the data be claimed?

No

5.6.2.8 Indicate whether data sharing will be postponed or restricted

Restricted

Personal data

5.6.2.9 Indicate who will be able to use the data

Nobody outside the CON-NET consortium unless they have an explicit written signed data sharing agreement with the consortium

5.6.2.10 Is it necessary to restrict access to certain communities or to apply a data sharing agreement?

Yes

5.6.2.11 Explain how and why access will be restricted to certain communities or data sharing agreements might apply

Data sharing agreement will be signed by legal representatives of all parties. Releasing the full raw data would in violation of GDPR. In the processed data, there is a risk of persecution, endangering the safety of individuals given that it might be used to categorise individuals as misinformation spreaders.

5.6.2.12 Explain what action will be taken to overcome or to minimise restrictions

A signing of a legally binding data sharing agreement

5.6.3 How will data for preservation be selected, and where data will be preserved long-term?

5.6.3.2 Indicate what data must be retained or destroyed for contractual, legal, or regulatory purposes.

Data sharing contract specifies that personal data should be destroyed after the end of the project unless its lifetime is extended.

5.6.3.3 Indicate how it will be decided what data to keep

Data can only be kept with written agreement with the project partner who collected the data. Data that is not personal data can always be kept.

5.6.3.4 Describe the data to be preserved long-term

Non-personal data related to the publications.

5.6.3.5 Explain the foreseeable research uses (and/ or users) for the data.

Analysis on online social media. For example, misbehaviour / misinformation.

5.6.3.6 Indicate where the data will be deposited

Zenodo (or similar platform), to be decided at a more practical date later in the project, once the data has advanced

5.6.3.7 Will there be established repository proposed?

No

5.6.3.8 Demonstrate that the data can be curated effectively beyond the lifetime of the grant

Not applicable

5.6.4 What methods or software tools are needed to access and use data?

5.6.4.1 Indicate how the data will be shared

A shared repository, supplementary materials, or other appropriate platforms

5.6.4.2 Indicate whether potential users need specific tools to access and (re-)use the data

Web browser

5.6.5 How will the application of a unique and persistent identifier to each data set be ensured?

5.6.5.2 Explain how the data might be re-used in other contexts

The data will be shared as part of publications, and might be used to confirm our results. Depending on the exact nature of data it might be used for other research purposes.

5.6.5.3 Indicate whether a persistent identifier for the data will be pursued

Yes

Digital Object Identifier (DOI)

Comment:

DOIs are given by standard in many repositories

6.7.2 Who will be responsible for data management?

6.7.2.2 Outline the roles and responsibilities for data management/stewardship activities

a. Mikko Kivelä (orcid:0000-0003-2049-1954)

Data management at Aalto University node of the project

b. Fintan McGee (orcid:0000-0001-7398-2664)

Data management at LIST node of the project

c. Onur Varol (orcid:0000-0002-3994-6106)

Data management at Sabanci node of the project

d. Michael Quayle (orcid:0000-0002-7497-0566)

Data management at University of Limerick node of the project

e. Nikolaos Deligiannis (orcid:0000-0001-9300-5860)

Data management at VUB node of the project

6.7.2.3 Is it a collaborative project?

Yes

6.7.2.4 Explain the co-ordination of data management responsibilities across partners

LIST will have the primary server for sharing data across the consortium. More details specified in the consortium agreement.

6.7.2.5 Indicate who is responsible for implementing the DMP, and for ensuring it is reviewed and, if necessary, revised

Fintan McGee (orcid:0000-0001-7398-2664)

6.7.3 What resources will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

6.7.3.2 Explain how the necessary resources (for example time) to prepare the data for sharing/preservation (data curation) have been costed in

The data management procedures are undertaken by the project personnel, under the supervision of the PIs. Good research data management practices and preparing data will take time from the researchers that is budgeted in the project, but using university servers and Zenodo do not incur extra costs.

6.7.3.3 Indicate whether additional resources will be needed to prepare data for deposit or to meet any charges from data repositories

No

Powered by

