Ethics guidance around the use of Online Datasets, Synthetic Datasets and Social Media Data

These notes should be considered when filling out the MACS Project Ethics Form or Infonetica Form.

Human data can involve username, user ID, name, address, personal details, and sensitive data among others. Note this list is not exhaustive.

When using human data from Online Datasets, Synthetic Datasets and/or Social Media Databases (e.g., Twitter (X), Meta (Instagram, Facebook), YouTube, etc) for research purposes you must adhere to the following guidelines.

The data you want to use can be classified either as non-publicly or publicly available.

Non-publicly available data

If the data is non-public, it is considered as private data.

1) Social media: if you have to log into the social media site to obtain data, then it is not considered to be public data. In this case you need to read the "terms of use" of the social media site on the use of private data of their users.
2) Non-public available datasets. You must also ensure you have authorization from the Datasource that you can produce.

Publicly Available Data

From Social Media

If you are gathering data without logging in and the data is available to the public, then you must:

- Preferably, not record user IDs and personal data of users (perform data anonymisation)
- If you record user IDs, you need to perform data pseudo-anonymisation
- Use only aggregated data (do not quote tweets or posts) when publishing and sharing
- Consider vulnerable users (e.g. children, political views from some countries)
- Take care if the data includes pictures or videos
- Check the terms and conditions of the social media site regarding use of the users' data
- Twitter (X) does not allow the scrapping of its services without their prior consent. Consider using the Twitter API

From Online Datasets

If you are downloading publicly available datasets from repositories or websites (e.g. Kaggle), you must ensure that the dataset has prior ethical clearance.

Examples of prior ethical clearance are:

- Ethical clearance from an ethics committee
- Data used in peer-reviewed articles, which shows prior ethical clearance

You must also ensure to perform data anonymisation.

Generating or Using Synthetic (made-up) Datasets

Some example generators:

- Faker Python package
- Mockaroo.com

Before using or creating a synthetic dataset please read this brief article: https://uksa.statisticsauthority.gov.uk/publication/ethical-considerations-relating-to-the-creation-and-use-of-synthetic-data/pages/6/

If you decide to generate or use a synthetic (made-up) dataset, you need to confirm that:

- The dataset is generated randomly
- The generator is not scrapping the internet to collect real data
- The data are used only for the intended purpose (e.g. your project).
- The dataset will NEVER be shared for other purposes such as other projects or made publicly available.