Plan Overview

A Data Management Plan created using DMPonline

Title: The Intersection of Human and Machine Similarity Models

Creator: Archibald Herbertson

Principal Investigator: Archie Herbertson

Data Manager: Archie Herbertson

Project Administrator: Archie Herbertson

Affiliation: University of Edinburgh

Template: UoE Default DMP template for PGRs

Project abstract:

Many recommender algorithms are trained to capture the complex relationships between involved items by reducing them to vector embeddings. The algorithm can retrieve a one number measure of the similarity between items by calculating the distance between their embeddings. Human field experts also use their understanding of complex similarity to make judgements about recommendations. By surveying these experts we can analyse how closely machine views of similarity match theirs. We can also investigate if models whose final recommendations are scored as more accurate by experts model similarity more similarly to the experts. Finally we will attempt to construct an embedding space directly from the data retrieved from the experts.

ID: 117241

Start date: 07-11-2022

End date: 28-04-2023

Last modified: 02-03-2023

The Intersection of Human and Machine Similarity Models

Administrative Information

1) School or Institute

- CSE - School of Informatics

2) Name and Contact details of supervisor(s)

Heather Yorston heather.yorston@ed.ac.uk
David Sterratt david.c.sterratt@ed.ac.uk

3) Project start date

2022-11-07

4) Project end date

2023-04-28

Data Collection

5) Data Collection

Data will be collected via surveys from up to 30 participants. Participants will respond to questions about how similar they see items from BBC data as being. Participants will only be shown BBC data that they already have access to.

Survey data will be collected from survey responders and collated via Microsoft Forms. This data will be saved to an Excel spreadsheet and then converted to a .csv file format.

Data produced by a range of recommender models will be used to build the survey questions. These "embeddings" extracted from the questionnaire will have been generated either by myself or the BBC who employ me as a degree apprentice. The similarity scores given by these embeddings that are used in the survey will be preserved in the survey PDF.

For the survey BBC articles will be referenced in the survey questions. They will be referenced by url, but will also be archived for usability of the survey data. The articles will be presented in the survey with their thumbnail, title and synopsis so this information will be preserved within a PDF. The number of articles involved will be less than 200.

Further exploration of the survey results will be done in a Jupyter Notebook which will include up to date executions of the code within. This file will be under 1GB and serve to present the analysis able to be performed using the survey results.

The University of Edinburgh offers courses on Data Protection and Data Protection for Research that

the PI will complete in order to ensure that the data is collected and managed using the highest standards and the best practices.

Documentation & Metadata

6) Documentation & Metadata

I will provide a database schema and index each file I submit within a README.txt. The ID metadata for survey responders will be provided but anonymised. I will record metadata for understanding the results of the survey in the README as well.. The attempt to generate embeddings from the results of the survey will be performed within a code notebook that is documented and commented for clarity. The results of this will also be labeled with relevant metadata.

Ethics & Legal Compliance

7) Ethics & Legal Compliance

There is no need to use the personal details of the survey responders in the analysis I will do, but a mailing list will be needed to follow up with them during the survey. As this is personal data it is necessary to hold it securely in DataStore and destroy it once it is no longer needed for the project. Emails made to this mailing list will address recipients in the BCC field so they are not shown who else is participating as this is personal data.

This data management plan will undergo Ethics review by the Informatics ethics and integrity board. A Data Protection Impact Assessment is included in the ethics review process so the continuation of this project will involve a successful Impact Assessment.

All data recorded in this project will be stored using the DataStore service. Copies of this data may be stored temporarily on a device protected to BBC standards which is only provided to employees who will already be authorised to process the data involved in this project. The device has a disk encrypted via FileVault. The device is password protected and, where possible, data will be processed on a user account that requires 2 factor authentication to sign into.

The methodology of this project and the BBC data used in it will go through review from the relevant legal department within the BBC to make sure it is legally compliant for the BBC. The released project will comply with the BBC's access and sharing regulations.

The participants will be informed before they enter the task that the provided data will be handled and stored securely, in line with GDPR best practice.

Storage and Back-Up

8) Where will your data be stored and backed-up during the project?

My data will be stored on MS Forms as the Outlook suite is officially adopted across the BBC, and

downloaded onto a BBC password protected laptop. This data will be uploaded onto DataStore for backup/recovery.

The backups provide resilience in the case of accidental deletion and against incidents affecting the main DataStore storage. The data are automatically replicated to an off-site disaster recovery facility, with 10 days of file history visible online. Off-site tape backups keep 60 days of history of the filesystem. The 60 day rolling snapshots allow important data to be recovered to a prior state, by request if beyond the visible period.

Sensitive data stored on DataStore will be further protected by the use of 256 bit encryption as required by University policy

Selection and Preservation

9) Where will the data be stored long-term?

Data will be stored for long term preservation in DataShare.

10) Which data will be retained long-term?

The embeddings for each BBC content item in the survey for each model involved in the study will be preserved in order to validate the results gained from comparing them to the survey results. The anonymised survey results will also be preserved for this purpose. There is not a need to preserve the articles referenced in the survey beyond the details shown about them within the survey PDF that will be preserved such as their URLs.

Data Sharing

11) Will the data produced from your project be made open?

- Yes: go to 12

12) How will you maximize data discoverability & access?

Anonymised and impersonal data will be made open using the DataShare repository after the completion of the project. No embargo period will be required and a persistent DOI will be created in the process of saving the data to DataShare.

Participant consent will be gathered through the participant information sheet. No confidentiality agreement will be required. The datasets will be stored in DataShare and shared under a CC BY 4.0 license. The decision on whether to supply data to a potential new user will be made by the PI after consultation with their line manager within the BBC.

Data will only be kept as exclusive during the completion of the project. Upon the completion

wherever the project is mentioned its DOI from DataShare will be promoted in order to increase its discoverability.

The anonymity of all participants supplying personal data should limit the risk of delays to sharing.

Responsibilities & Resources

14) Who will be responsible for the research data management of this project?

Responsibilities - Archie Herbertson. Seek help from Informatics Data Manager to ensure best practice in Data management is followed.

Governance of access - The decision on whether to supply the data to a potential new user will be made by the PI after consulting the BBC.

The data will be distributed under a CC BY 4.0 license.

15) Will you require any training or resources to properly manage your research data throughout this project?

This project necessitated the completion of the Data Protection Training and Data Protection For Research courses in LEARN. DataShare is free and the operations involved in completion of the coding aspects of the project will all be completable within the resources I have access to as an apprentice with the BBC.