Plan Overview

A Data Management Plan created using DMPonline

Title: Conversational Agents for Ethical Concerns in Software Evolution

Creator: Lauren Olson

Data Manager: Lauren Olson

Affiliation: Vrije Universiteit Amsterdam

Funder: European Research Council (ERC)

Template: ERC DMP

Project abstract:

To create and maintain software that is useful, usable and ethical, user needs, expectations and concerns should be considered in the pre and post deployment phases. Users increasingly express feedback about software applications through social media and specialized user feedback platforms, which may be automatically identified and summarized to reduce manual workload.
In this project we will address two challenges connected to this automated extraction of user feedback:
The first is to ensure the actionability of user feedback. We will do this by detecting whether any relevant aspects of user feedback are missing and deploying a conversational agent to ask the user for missing aspects, if any.
The second challenge is to elicit ethical concerns related to the software from end-users that give feedback on social media. We will do this by developing a conversational agent that asks users from diverse backgrounds about ethical concerns that should be addressed in the software application.

ID: 115554

Start date: 01-11-2021

End date: 01-11-2025

Last modified: 24-01-2023

Copyright information:

Conversational Agents for Ethical Concerns in Software Evolution

Summary

Project Acronym

CEF

Project Number

HRC0034958

Provide a dataset summary

For this project, we use user feedback from multiple channels (social media, app stores) to determine users' software requirements. This data is collected via web scraping, typically from existing web scrapers from Github. These datasets need to be updated on a regular basis (within a year or so) in order to accurately represent users' interests. As such, data is typically collected on a large scale from a variety of social media populations or App Store applications. The size of this data depends on the targeted applications and user populations but typically includes a few hundred thousand to a few million data entries. Each entry contains a unique identifier, time stamp, content, evaluation metrics, and username, typically collected as a string or float datatype.

FAIR data and resources

1. Making data findable

When each research paper is published, as is typical within the software engineering community, I will a replication package available on Zenodo. Zenodo follows FAIR data principles for their site, requiring users to fill in relevant metadata to make their work findable to different communities before posting.

This replication package includes all code I use to generate the datasets, manipulate the data, and produce the reported results within the work. It also includes a README.md file which contains information relating to the environment on which I ran the programs. For example, I have a 2021 MacBook Pro and run all my programs with Python 3.9. The README.md also includes a short description of the project, each file in the repository, and any references I used to write the code. Finally, it includes instructions for running the programs in order to reproduce the results. Next, I also include a requirements.txt file which allows fellow researchers to easily download all the other software packages I use within my code and provides a source for researchers to inspect these resources.

Although I provide links to the code I use to collect data, because I collect online user-generated content, my data collection is temporally dependent. As such, I provide information and proof of the

time my data is collected as well as information in the Limitations sections of my research papers explaining how this may affect the generalisability of the data.

2. Making data openly accessible

Once my research is accepted to be published, I will not only have my data available on Zenodo, which is Open Access, but will publish these datasets to Github as well. My Github account is linked to the VU's Software and Sustainability group, where we keep all replication packages. This allows the data to be easily accessible from a diverse group of researchers with an interest in the VU's S2 group. We also make an effort to advertise our work online on various outlets to maximise outside access.

3. Making data interoperable

Within my python code, I use all standard style conventions (https://peps.python.org/pep-0008/) to ensure that code is readable. This includes organization and naming standards for data and variables. When web scrapers are used to collect data, typically this data is stored in a JSON file format which stores each data entry as an array of dictionaries. These dictionaries includes the datatype with a reference to the specific instance of data. These JSON files, due to their structure, are easily translatable to a CSV file format, where datatypes become the file headers and then the instances become the rows within the CSV files. This data can then easily be manipulated through python packages like pandas, which stores CSV files in data frames for manipulation. This data is then posted to Github and is thus available for other researchers to easily use.

4. Increase data reuse

My data will be completely available for other researchers to use. All my data and code is posted online to maximise reproduction of results. This data is posted to both Zenodo and Github, with references to these links in the research paper itself as well as on official VU sites and social media to increase visibility of the work.

My data can be reused for replication as well as different types of quantitative and qualitative analysis at varying levels. My code could be reused on other similar data to facilitate this process, especially the machine learning explainability features, which will hopefully provide other researchers insight into how their models work.

5. Allocation of resources and data security

Although we use user-generated data, we do not use human participants within our research. All resources we use to collect and store data, at this point of the project, are free. Our research is secure as we use non-local machines to store our data. The multiple cloud services we use store our data in datacenters which have state of the art security systems and longevity requirements.