Goal-setting dialogue for physical activity with a virtual coach.

Author: Beyza Hizli

This repository contains the data and code related to the final experiment of the thesis project *Goal-setting dialogue for physical activity with a virtual coach*.

For a comprehensive discussion about the results, please refer to Chapter 4 of the thesis: . For a more detailed description of the final experiment, please refer to the OSF registration: https://osf.io/4duwh

Contents of the repository

File/Folder name	Content
Dockerfile	File containing instructions to build Docker image
README.md	(This file) containing instructions and information on the data in this folder
README.pdf	PDF file of this file
bayesian_analysis/	Folder containing R code for the bayesian analysis and the corresponding data
- bayesian_analysis.Rmd	R Markdown file containing the Bayesian analysis
- bayesian_analysis.pdf	PDF file that results from knitting the bayesian_analysis.Rmd file
- bayesian_analysis_df.xlsx	Preprocessing of analysis data to create the dataframes including explanations of the variables
- df_acceptance.csv	CSV file containing acceptance ratings
- df_ratings.csv	CSV file containing motivation ratings
- df_se_plots.csv	CSV file containing running and walking self-efficacy scores, transformed
- difference_dataframes.ipynb	Preprocessing data to analyze difference between the two groups
- participant_characteristics.Rmd	R Markdown file to output participant characteristics table
- participant_characteristics.pdf	PDF file that results from knitting the participant_characteristics.Rmd file
- pre_C.csv	Participant data of final experiment
- smoker.csv	Number of people per group per condition for the smoke variable
- ttm_phase.csv	Number of people per group per condition for the ttm phase for physical activity variable

File/Folder name	Content
- weekly_exercise.csv	Number of people per group per condition for the weekly exercise variable
data/	Folder containing the data retrieved from Prolific and Qualtrics
- examples.xlsx	Examples of people that achieved a running or walking goal, including an introduction of the person, the goal they achieved and how they achieved it
- post_questionnaire_C.xlsx	Post-questionnaire data from the Qualtrics questionnaire
- pre_questionnaire_C.xlsx	Pre-questionnaire data from the Qualtrics questionnaire
- prolific_data_AB.xlsx	User data retrieved from the participants' Prolific profiles from the data collection experiments
- prolific_data_C.xlsx	User data retrieved from the participants' Prolific profiles from the final experiment
- questionnaire_data_AB.xlsx	Questionnaire data from the Qualtrics questionnaire of the data collection experiments
- reading_time.xlsx	Reading times retrieved from the conversations with the virtual coach.
data/chats	Folder containing data from the chats with the virtual coach. The file name represents the anonymized participant id.
model/	Folder containing files for the creation of the prediction model
- AB_data.csv	User data of the participants from the data collection experiment
- AB_data_preprocessing.xlsx	Initial preprocessing file of the participant data from the data collection experiment
- B_ratings	Motivation ratings that the participants of the second data collection experiment have given to the examples
- C_final	Cleaned user data of the participants from the data collection experiment
- C_prediction	CSV file containing the motivation rating predictions that the final prediction model made
- centers.csv	CSV file containing the cluster centers
- clustering.Rmd	R Markdown file containing the clustering algorithm
- clusters.csv	CSV file containing the clusters
- collaborative_filtering.ipynb	Notebook containing the collaborative filtering algorithm
- Dockerfile	Dockerfile to run a Jupyter Lab container
- km_clustered.png	Image of the clusters

File/Folder name	Content
- model_analysis.Rmd	R Markdown file for the analysis of the prediction model
- motivation_final.csv	CSV file containing the input variables for the motivation rating prediction model
- participant_characteristics_AB.Rmd	R Markdown file to output participant characteristics table for parts A and B
- participant_characteristics_AB.pdf	PDF file that results from knitting the participant_characteristics_AB.Rmd file
- prediction_model.Rmd	R Markdown file containing the final prediction model
- preprocessing_AB.ipynb	Preprocessing code for the participant data of the data collection experiments
- similarity_final.csv	CSV file containing the input variables for the similarity rating prediction model
- similarity_predictions.csv	CSV file containing the similarity predictions of the collaborative filtering algorithm
predictions_final_experiment/	
- C_data_preprocessing.xlsx	Initial preprocessing file of the participant data from the final experiment
- C_final.csv	Cleaned user data of the participants from the final experiment
- C_prediction.csv	CSV file containing the motivation rating predictions that the final prediction model made
- Dockerfile	Dockerfile to run a Jupyter Lab container
- pre_AC	User data of the participants from the final experiment (C) and user data of the example people (A)
- predictions_C.ipynb	Notebook that contains code to extract top three predicted examples for the participants of the final experiment
- prepocessing_C.ipynb	Preprocessing code for the participant data of final experiment
thematic_analysis/	Folder containing files for the thematic analysis
- coder1.csv	Codes of the first coder. These codes are taken from the thematic_analysis_full.xlsx file.
- coder2.csv	Codes of the second coder. These codes are taken from the thematic_analysis_full.xlsx file.
- Dockerfile	Dockerfile to run a Jupyter Lab container
- thematic_analysis.ipynb	Notebook to get the inter-rater reliability score of the coders

File/Folder name	Content
- thematic_analysis_full.xlsx	File containing the messages that are coded with their codes by the two coders

Initial Setup

Pre-requisites:

- Docker (You can follow the instructions here)

Obtaining the Docker image

You can build the Docker image in one of the following two ways:

- Pull the Docker image from Docker Hub using the following command:

```
docker pull bhizli/thesis_analysis_goal_setting_virtual_coach
```

OR

- Build the image from the given Dockerfile
- - Ensure Docker daemon is running
- - Open Terminal or Command Prompt
- - Run the following command:

```
docker build -f /path/to/a/Dockerfile -t <DOCKER_IMAGE_NAME> .
```

More information regarding `Dockerfile` and how to build images can be found here

Running container RStudio

Use the following command to run a Docker container from the Docker image you have built from the previous step.

```
docker run -d -p 8787:8787 -v <PATH_TO_CURRENT_FOLDER>:/home/rstudio/analysis -e
PASSWORD=<PASSWORD> <DOCKER_IMAGE_NAME>
```

If you are running a container from the image pulled from Docker Hub, the command should as follows:

```
docker run -d -p 8787:8787 -v <PATH_TO_CURRENT_FOLDER>:/home/rstudio/analysis -e
PASSWORD=<PASSWORD> bhizli/thesis_analysis_goal_setting_virtual_coach
```

- If the container has started successfully, you should see a new container in Docker Desktop under `Containers/Apps`.
- Navigate to `https://localhost:8787`, and you should see a login page for RStudio. Enter `rstudio` as the username, and the password you specified when you ran the container to login.
- RStudio should be running in your browser. You can start running the code now.

Running .ipynb files

We explain two ways:

Jupyter Lab

You can use Jupyter Lab to run .ipynb files

To install Jupyter Lab, first make sure to have Python 3 and pip installed on your device. Visit https://www.python.org/ to download the latest version of Python (versions later than 3.4 include pip by default).

You can check whether you have Python and/or pip and their versions by running the following commands: `python -version` and `pip -version`.

To install Jupyter Lab, run the following command: `pip install jupyterlab` To run .ipynb files, navigate to the corresponding folder and run the following command: `jupyter lab`

This will open Jupyter Lab in your default browser. You can run the code by navigating to the 'Run' tab on the top bar and selecting 'Run All Cells'.

Running a separate Docker container

Alternatively, you could build the docker image from the Dockerfile that is located in the folder with the .ipynb files.

```
docker build . -t <image_name>
```

After building the image, you can run the following command to access Jupyter Lab in the browser:

```
docker run -p 8888:8888 -e JUPYTER_ENABLE_LAB=yes -v
<this_working_directory>:/home/jovyan/work <image_name>
```

where <this_working_directory> should be the directory where the .ipynb files are located.

Knitting R Markdown file

- Navigate to `https://localhost:8787`, and you should see a login page for RStudio. Enter `rstudio` as the username, and the password you specified when you ran the container to login.

- To knit the .Rmd files, press "Knit" on the toolbar and select "Knit to PDF". The PDF will be saved to the folder the .Rmd file is located in.

Reproducing the same results

To provide a better understanding of how the complete analysis is conducted and in which order, we give an overview below.

Preprocessing data for prediction model input

Chapter 3, Section 3.3-3.5

1. We start with raw participant data:
    - `data/prolific_data_AB.xlsx`
    - `data/questionnaire_data_AB.xlsx`
2. To process the data in these raw files to obtain the user variables, the two files are combined in one .xlsx file and the averages are taken where needed. This is done in:
    - `model/preprocessing_AB.xlsx` The relevant columns, which are the resulting user variables that are used for the model, are extracted and can be found in one .csv file:
    - `model/AB_data.csv`
3. In addition to the user variables, we use the cluster variables as input for the prediction models. For that, we first need the similarity predictions that are computed in the following file:
    - `model/collaborative_filtering.ipynb` This outputs:
    - `model/similarity_predictions.csv`
4. These similarity predictions are used by the clustering algorithm in the following Rmd file:
    - `model/clustering.Rmd` This outputs the clusters and cluster centers:
    - `model/clusters.csv`
    - `model/centers.csv`
5. To finalize the creation of the cluster variables and combine them with the user variables, we run the following .ipynb file:
    - `model/preprocessing_AB.ipynb` This outputs the following two files that will be used as input for the model analysis:
    - `model/similarity_final.csv`
    - `model/motivation_final.csv`

Prediction model analysis

Chapter 3, Section 3.6

The analysis of the prediction model can be found in:

- `model/model_analysis.Rmd`

The best performing model can be found separately in:

- `model/prediction_model.Rmd` This model is used to make motivation rating predictions.

Preprocessing data to make predictions for the final experiment

Part of preparation for the experiment in Chapter 4

1. We start with raw participant data:
    - `data/prolific_data_C.xlsx`
    - `data/pre_questionnaire_data_C.xlsx`
2. To process the data in these raw files to obtain the user variables necessary for the prediction model, the two files are combined in one .xlsx file and the averages are taken where needed. This is done in:
    - `predictions_final_experiment/C_data_preprocessing.xlsx` The relevant columns are extracted and put into: `predictions_final_experiment/pre_C.csv` A combination of the user data of the participants in the final experiment and the participants that provided examples can be found in: `predictions_final_experiment/pre_AC.csv`
3. To finalize the user variables and create an input file for the prediction model, we run the following .ipynb file:
    - `predictions_final_experiment/preprocessing_C.ipynb` This outputs the following two files that will be used as input for the model analysis:
    - `predictions_final_experiment/C_final.csv`
4. This file is used by the prediciton model to make predictions:
    - `model/prediction_model.Rmd` Which results into:
    - `model/C_prediction.csv`
5. To extract the top three predicted examples per participant, we run:
    - `predictions_final_experiment/predictions_C.ipynb` Which shows the predictions for the participants.

Preprocessing data for the final analysis

Chapter 4, Section 4.1-4.2

1. We start with raw analysis data:
    - `data/post_questionnaire_data_C.xlsx`
2. To process the data in this raw files to obtain the variables necessary for the Bayesian analysis, the averages are taken where needed. This is done in:
    - `bayesian_analysis/final_analysis_data_preprocessing.xlsx` from this sheet, we can extract the following dataframes:
    - `bayesian_analysis/df_acceptance.csv`
    - `bayesian_analysis/df_ratings.csv`
    - `bayesian_analysis/df_se.csv`
    - `bayesian_analysis/df_se_plots.csv`
3. To run the Bayesian analysis, open the following Rmd file:
    - `bayesian_analysis.Rmd` In this file you can find the results of the final analysis.

Thematic analysis

Chapter 4, Section 4.2.1

To obtain the inter-rater reliability score of the coding for the thematic analysis, run:

- `thematic_analysis/thematic_analysis.ipynb`

Other

To obtain the participant characteristics table, run:

- `bayesian_analysis/participant_characteristics.Rmd`

Tables/Figures and corresponding files

File name	**Corresponding figure/table in the report**
prediction_model.Rmd	Table 3.2: Final model used to predict motivation rating
participant_characteristics.Rmd	Table 4.1: Participant characteristics.
bayesian_analysis.Rmd	Table 4.2: Differences between the two groups (General and Personalized group).
bayesian_analysis.Rmd	Figure 4.2: Comparison self-efficacy before (pre) and after (post) the conversation with the virtual coach.
bayesian_analysis.Rmd	Table 4.3: Mean, standard deviation and mean difference of the selfefficacy of the two groups (general examples and personalized examples).
bayesian_analysis.Rmd	Figure 4.3: Comparison selfefficacy before (pre) and after (post) the conversation with the virtual coach per group (general or personalized examples).
bayesian_analysis.Rmd	Table 4.4: Mean, standard deviation and mean difference of the motivation ratings for general and personalized examples.
bayesian_analysis.Rmd	Figure 4.4: Comparison of motivation ratings between the two example types: General and Personalized.
bayesian_analysis.Rmd	Table 4.5: Means and standard deviations for each item of the acceptance questionnaire. 'Average' shows the mean and standard deviation of all 6 questions combined (overall acceptance score).
bayesian_analysis.Rmd	Figure 4.5: Box plots of each question and overall score of the acceptance questionnaire.
participant_characteristics_AB	Table B.1: Participant characteristics part A.
participant_characteristics_AB	Table B.2: Participant characteristics part B.

File name	Corresponding figure/table in the report
model_analysis.Rmd	Table F.1: Summary of model predicting similarity rating with all independent variables.
model_analysis.Rmd	Table F.2: Summary of model predicting motivation rating with all independent variables.
model_analysis.Rmd	Table G.1: Model using all available variables to predict motivation rating.
model_analysis.Rmd	Table G.2: Stepwise regression model predicting motivation rating
model_analysis.Rmd	Table G.3: Correlation between the independent variables and dependent variable (motivation rating).