Step 5: Training the Model & Results

Author: Nele Albers

Date: December 2024

This README-file is meant to give you an overview of the data and analysis code underlying our 5th pipeline step and results presented in the "Results"-section.

Authored by Nele Albers, Mark A. Neerincx, and Willem-Paul Brinkman.

Types of analyses

Below we describe how you can reproduce/inspect our results for the different types of analyses that we have conducted for step 5 of our pipeline and our results presented in the "Results"-section.

Anonymization and data

The files "anonymize_sessiondata.py," "clean_and_anonymize_prescreening_data.py," "clean_and_anonymize_postquestionnaire_data.py," and "clean_and_anonymize_followupquestionnaire_data.py" in the "Data"-folder show how we anonymized the data from the questionnaires and conversational sessions. Since we cannot provide the raw, non-anonymized data, you cannot run these files yourself. For the post-questionnaire and follow-up questionnaire, there are also output files that provide information such as the number of submissions removed for failing too many attention checks.

The resulting data files are "prescreening_questionnaire_anonym.csv," "prolific_profile_anonym.csv," "sessionsdata_anonym.csv," "postquestionnaire_anonym.csv," and "followupquestionnaire_anonym.csv" in the "Data"-folder. For each data file, there is an .xlsx-file that explains each data column.

Preprocessing

The file "merge_data.py" merges the anonymized data from the prescreening questionnaire, Prolific profiles, and conversational sessions and creates the file "data_rl_samples_binary.csv" with the RL samples that underlies all following RL-based analyses. The file "merge_data_output.txt" contains some log data from this process, such as the feature means that were used for creating binary user-inquired state features.

The file "process_data_for_time_comparison.py" processes the data from the prescreening questionnaire, post-questionnaire and follow-up questionnaire to create the data files needed for comparing the smoking frequency, weekly exercise, and quitter self-identity between the prescreening questionnaire on the one hand and the post-questionnaire and follow-up questionnaire on the other hand. Running the file also produces the file "process_data_for_time_comparison_output.txt" that shows the Cronbach's alpha values we computed for quitter self-identity.

Comparing smoking frequency, weekly exercise, and quitter self-identity over time

Navigate to the folder "Analysis_Time_Comparison" and follow the instructions in the README-file therein to reproduce:

- our comparison of smoking frequency, weekly exercise, and quitter self-identity between the prescreening questionnaire and post-questionnaire reported in "Step 5: Training the model - Study", and
- our comparison of smoking frequency, weekly exercise, and quitter self-identity between the prescreening questionnaire and follow-up questionnaire reported in "Step 5: Training the model - Study."

Data overview, participant characteristics, feature selection, and AQ1

Navigate to the folder "DataOverview_ParticipantCharacteristics_FeatureSelection_AQ1" and follow the instructions in the README-file therein to reproduce:

- our data overview and participant characteristics:
  - the mean effort per activity (cluster) from the "Step 5: Training the model"-section and Table S10 in the Appendix,
  - the mean effort per preparatory activity cluster and combination of values for the three selected user-inquired features from Figure S3 in the Appendix,
  - the number of samples per activity (cluster) and combination of values for the three selected user-inquired features from Figure S4 in the Appendix,
  - the dropout response per session from the "Step 5: Training the model"-section, and
  - the participant characteristics (e.g., age, gender, smoking frequency) from Table S8 in the Appendix.
- our selection of three user-inquired state features based on our collected data.
- our analysis for AQ1:
  - Figure 5.4 from the chapter,
  - the Cohen's h values from our results for AQ1,
  - Figure S5 from the Appendix, and
  - the examples of optimal activities in the eight possible starting states from our setup for AQ1.

AQ2: Changing usefulness beliefs

Navigate to the folder "AQ2_Changing_Usefulness_Beliefs" and follow the instructions in the README-file therein to reproduce our analyses for our second analysis question.

Explanation of files and folders

This directory contains the following files and folders:

- Analysis_Time_Comparison: To reproduce our comparison of smoking frequency, weekly exercise, and quitter self-identity between the prescreening questionnaire on the one hand and the post-questionnaire and follow-up questionnaire on the other hand.
- AQ2: To reproduce our analyses for our second analysis question.
- Data:
  - anonymize_sessionsdata.py: Code used to anonymize the data from the conversational sessions.
  - clean_and_anonymize_followupquestionnaire_data.py: Code used to anonymize the data from the follow-up questionnaire.
  - clean_and_anonymize_followupquestionnaire_data_output.txt: Output of running the above file (e.g., number of submissions removed for failing

too many attention checks).
    - clean_and_anonymize_postquestionnaire_data.py: Code used to anonymize the data from the post-questionnaire.
    - clean_and_anonymize_postquestionnaire_data_output.txt: Output of running the above file (e.g., number of submissions removed for failing too many attention checks).
    - clean_and_anonymize_prescreening_data.py: Code used to anonymize the data from the Prolific profiles and prescreening questionnaire.
    - followupquestionnaire_anonym.csv: Anonymized data from the follow-up questionnaire (e.g., smoking frequency).
    - followupquestionnaire_data_explanation.xlsx: Explanation of the data columns in the file above.
    - postquestionnaire_anonym.csv: Anonymized data from the post-questionnaire (e.g., smoking frequency, weekly exercise amount).
    - postquestionnaire_data_explanation.xlsx: Explanation of the data columns in the file above.
    - prescreening_questionnaire_anonym.csv: Anonymized data from the prescreening questionnaire (e.g., smoking frequency).
    - prescreening_questionnaire_data_explanation.xlsx: Explanation of the data columns in the file above.
    - prolific_profile_anonym.csv: Anonymized Prolific profile data of participants.
    - prolific_profile_data_explanation.xlsx: Explanation of the data columns in the file above.
    - sessionsdata_anonym.csv: Anonymized data from the conversational sessions with the virtual coach Mel.
    - sessionsdata_data_explanation.xlsx: Explanation of the data columns in the file above.
- DataOverview_ParticipantCharacteristics_FeatureSelection_AQ1: To reproduce our data overview, participant characteristics, feature selection, and analyses for AQ1.
- merge_data.py: Code to merge the anonymized data from the prescreening questionnaire, Prolific profiles, and conversational sessions and create RL samples.
- merge_data_output.py: Output of the above file (e.g., means of user-inquired state features).
- process_data_for_time_comparison.py: Code to process the data from the prescreening questionnaire, post-questionnaire, and follow-up questionnaire for the comparison of smoking frequency, weekly exercise, and quitter self-identity over time.
- process_data_for_time_comparison_output.txt: Cronbach's alpha values we computed for quitter self-identity as output of running the file above.
- README.md/README.pdf: This README-file.