# **Code**

# main.ipynb The only file for all experiments in the paper, the 3 main experiments and the 2 additional analyses. The visualization code is also provided in this file.

# **Data**

* responses.csv  
  All participants responded to all 100 questions in the IPIP Big5 questionnaire using a custom application. The responses have integer values from 1 to 5. Responses have been reverse coded according to their direction on the respective psychological construct. The final dataset has a dimension of 1000 x 100 with each row standing for a participant and each column for a question.
* users\_embeddings\_BERT.csv  
  Containing BERT embeddings as features for users. The embeddings have been acquired and pre-processed using the following methods:
  + Collected posts of 1000 Facebook users, under 65, with at least 300 posts and1000 words
    - Randomly picked 300 posts
    - Example post: "Someone spoiled my good mood... :("
  + Used a pre-trained BERT Model from HuggingFace (<https://huggingface.co/transformers/model_doc/bert.html>) to extract word embeddings
    - BERT Base Uncased (12 layers, 768 dimensions)
    - Averaged last four layers to get the word embeddings
    - Averaged embeddings across words to get the post embeddings
    - Averaged post embeddings to get the user embeddings
  + Processing steps:
    - Standardized using Sklearn’s StandardScaler()
    - PCA to 90% Variance (N = 250)
  + Final Dataset dimensions: (1000 x 250), with 1 header row
  + Rows: Participants, Columns: Embeddings features
* users\_embeddings\_LDA.csv  
  Containing LDA as features for users.
  + The set of users and posts are similar to users\_embeddings\_LDA.csv
  + Features are extracted from a pre-trained, publicly available LDA topics at <https://dlatk.wwbp.org/datasets.html?highlight=met_a30_2000_cp>, using the DLATK library available at <https://dlatk.wwbp.org/index.html>
  + Processing steps:
    - Standardized using Sklearn’s StandardScaler()
    - PCA to 90% Variance (N = 358)
  + Final Dataset dimensions: (1000 x 358), with 1 header row
  + Rows: Participants, Columns: Embeddings features
* questions\_embeddings.csv  
  The embeddings of question texts from the IPIP Big 5 questionnaire
  + Used a pre-trained BERT Model to extract word embeddings
    - BERT Large and Uncased (24 layers, 1024 dimensions)
    - Averaged last four layers to get the word embeddings
    - Averaged word embeddings to get the respective question’s embeddings
  + Processing steps:
    - Standardized using Sklearn’s StandardScaler()
    - PCA to 90% Variance (N = 55)
  + Final Dataset dimensions: (100 x 55), with 1 header row
  + Rows: Question, Columns: Embeddings features
* construct\_list.csv  
  File containing the construct (category) and key (direction - negative or positive) of each question in the questionnaire. This file also contains the question texts shown to the participants.

# 