# READ ME FILE

**LILLA BÍBORKA DÖMSÖDI**

*Spring, 2024*

This GitHub Repository houses the scripts utilized in conducting the study forming the basis of the Master's Thesis titled "Exploring the Role of Online Social Support in Eating Disorder Recovery: A Theory-Driven Computational Study." The Thesis was crafted as part of the MSc program in Social Scientific Data Analysis at Graduate School, Faculty of Social Sciences, Lund University.

The original datasets are not provided within this repository for ethical research considerations. However, fellow social scientists and interested parties are encouraged to replicate the study using their own manually collected user base. For any inquiries, please do feel free to contact the author at domsodilbiborka@gmail.com.

Here, you'll find a comprehensive description of each script's environment, purpose, and the necessary input files, as well as the output results. At the time of writing, running these code cells is completely free of charge and accessible to everyone. However, please note that memory usage and software requirements may vary. Additionally, certain scripts may require extended running times, potentially spanning multiple hours, especially when dealing with large user bases.

# 1_manual_user_collection.Rmd

**Environment:** R

**Aim:** Processing the manual username collection, extracting the information for the API Data extraction.

**Input File(s):** Manual Username Collection in xlsx format, containing the below columns

| COLUMN NAME | label | user_name | submission_id | comment_link |
|---|---|---|---|---|
| VALUE | recovered / non-recovered | Reddit username of the detected user; must be pasted into the cell without any typos or extra spaces | submission_id of the post where the given user declared their recovery status; Blank if they expressed it in a comment | Comment_url of the comment where the given user declared their recovery status; Blank if they expressed it in a submission |

**Output File(s):** 1.1_usernames.txt, 1.2_submissions.txt, 1.3_comments.txt

---

# 2_api_data_collection.ipynb

**Environment:** Python [Google Colab]

**Aim:** Pulling the user history datasets from the Reddit API.

**Input File(s):** 1.1_usernames.txt, 1.2_submissions.txt, 1.3_comments.txt

**Output File(s):** 2.1_submissions_data.xlsx, 2.2_comments_data.xlsx

---

## 3_data_cleaning.Rmd

**Environment:** R

**Aim:** Conducting data inspection, cleaning, and verification steps.

**Input File(s):** 2.1_submissions_data.xlsx, 2.2_comments_data.xlsx

**Output File(s):** 3.1_data_cleaning_output.RData

---

## 4.1_analysis1.Rmd

**Environment:** R

**Aim:** Computing sub-indicators on the user level for the Social Integration component: number of interactions, number of connections, temporal frequency of interactions.

**Input File(s):** 3.1_data_cleaning_output.RData

**Output File(s):** 4.1.1_analysis1_output.RData

---

## 4.2_analysis2.Rmd

**Environment:** R

**Aim:** Computing sub-indicators on the user level for the Social Network Structure component: reciprocity, ego-network density, and eigenvector centrality.

**Input File(s):** 4.1.1_analysis1_output.RData

**Output File(s):** 4.2.1_analysis2_output.RData

---

## 4.3.1_sentiment_analysis.ipynb

**Environment:** Python [Google Colab]

**Aim:** Test the performance of 4 different pre-trained sentiment analysis models. Then, run the selected model on the comments dataset and attach the corresponding sentiment score to the original dataset.

**Input File(s):** 2.2_comments_data.xlsx

**Output File(s):** 4.3.1.1_comments_data_sa.xlsx

---

## 4.3.2_analysis3.Rmd

**Environment:** R

**Aim:** Computing the indicator on the user level for the Relational Content component: average sentiment score. Cleaning the dataset to only include the final, analysis-ready table.

**Input File(s):** 4.2.1_analysis2_output.RData, 4.3.1.1_comments_data_sa.xlsx

**Output File(s):** 4.3.2.1_analysis3_output.RData

---

## 5_variable_inspection.Rmd

**Environment:** R

**Aim:** Inspecting the distribution of the sub-indicators, followed by the creation of compound variables for Social Integration, Social Network Structure, and Social Support. Visualization of binary relationships between these variables and Recovery. Inspecting the linear relationship between the predictor variables.

**Input File(s):** 4.3.2.1_analysis3_output.RData

**Output File(s):** 5.4_variable_inspection_output.RData

*Computed Variable Distributions:*

5.1.0_interactions.jpg, 5.1.1_connections.jpg, 5.1.2_frequency.jpg,
5.1.3_social_integration_index.jpg, 5.1.4_reciprocity.jpg,

5.1.5_ego_network_density.jpg, 5.1.6_eigenvector_centrality.jpg, 5.1.7_social_network_structure_index.jpg, 5.1.8_relational_content.jpg, 5.1.9_social_support_index.jpg, 5.1.10_recovery_label.jpg

*Bivariate Relationship Plots:*

5.2.1_bivariate1.jpg, 5.2.2_bivariate2.jpg, 5.2.3_bivariate3.jpg, 5.2.4_bivariate4.jpg, 5.2.5_bivariate5.jpg

*Interaction Plots:*

5.3.1_interaction1.jpg, 5.3.2_interaction2.jpg, 5.3.3_interaction3.jpg, 5.3.4_interaction4.jpg

---

# 6_model_building.Rmd

**Environment:** R

**Aim:** Building 3 logistic regression analysis models: Baseline Model, Elaboration Model, Exclusionary Model. Testing Assumptions. Plotting calculated Probabilities.

**Input File(s):** 5.4_variable_inspection_output.RData

**Output File(s):**

*Model Summary Tables:*

6.1.1_model1.htm, 6.1.2_model2.htm, 6.1.3_model3.htm

*Probability Plots:*

6.2.1_probability1.jpg, 6.2.2_probability2.jpg, 6.2.3_probability3.jpg