MASTER'S THESIS 2025

Predicting Psychological Scale Scores and Deltas from Structured Word Responses: A Comparative Study of Regression Pipelines

Marcel Urrutia Nilsson

ISSN 1650-2884
LU-CS-EX: 2025-21

DEPARTMENT OF COMPUTER SCIENCE
LTH | LUND UNIVERSITY

EXAMENSARBETE
Datavetenskap

LU-CS-EX: 2025-21

Predicting Psychological Scale Scores and Deltas from Structured Word Responses: A Comparative Study of Regression Pipelines

Att förutsäga poäng på psykologiska formulär och deras förändringar från strukturerade ordsvar: En jämförande studie av regressionspipelines

Marcel Urrutia Nilsson

Predicting Psychological Scale Scores and Deltas from Structured Word Responses: A Comparative Study of Regression Pipelines

Marcel Urrutia Nilsson
marcelunilsson@gmail.com

June 11, 2025

Master's thesis work carried out at
the Department of Computer Science, Lund University.

Supervisor: Dennis Medved, dennis.medved@med.lu.se
Examiner: Maj Stenmark, maj.stenmark@cs.lth.se

Abstract

Background: Despite the potential benefits of using structured word responses for predicting psychological construct using rating scale scores, there is a lack of research in this area, particularly in the context of construct scale delta prediction. Thus, a comparative study of regression pipelines could provide valuable insights into the most effective method for this type of analysis.

Aim: This thesis aims to investigate the effectiveness of various regression pipelines in predicting psychological construct scale scores and deltas using structured word responses.

Method: The data used is from an earlier study with 477 participants who completed the Harmony in Life Scale (HILS) and the Satisfaction With Life Scale (SWLS) twice. Word embedding extraction and PCA were used for preprocessing, and two feature engineering strategies were employed for delta prediction: Delta vector and T1-score. Hyperparameters were optimized using Bayesian optimization, and four regression models were compared: ridge regression, Bayesian ridge regression, random forest, and AutoKeras StructuredDataRegressor.

Results: AutoKeras outperformed all other models for both score and delta prediction, with Pearson r-values of 0.929 and 0.918 for the HILS and SWLS score prediction, and 0.933 and 0.874 for the HILS and SWLS delta prediction. No other model achieved an r-value higher than 0.8 in any task.

Conclusions: The findings suggest that AutoKeras is a promising approach for predicting psychological construct scale scores and deltas from structured word responses. However, several modifications are necessary to ensure the validity and comparability of the results, including addressing data leakage and adopting a format more consistent with earlier studies. Further research and bigger datasets are needed to evaluate the generalizability of these results.

Keywords: psychology, NLP, regression, transformer, word embeddings, AutoKeras

Contents

0.1 Introduction

In psychology, a core challenge is how to reliably measure abstract mental states such as well-being, depression, or satisfaction. These abstract ideas are called *psychological constructs*, and they are often evaluated using closed-ended rating scales. However, these scales can over-simplify complex, multidimensional phenomena.

This thesis explores whether structured free-text responses—ten words participants use to describe how they feel—can be used to predict psychological construct scale scores, and more interestingly, the change in those scores over time. The input to the models is therefore natural language data from these word responses, and the output is either the predicted rating scale score (e.g., on the Satisfaction With Life Scale) or the *delta*, i.e., the change in score between two time points (T1 and T2).

To evaluate well-being, this thesis focuses on two validated psychological rating scales: the Satisfaction With Life Scale (SWLS) and the Harmony in Life Scale (HILS). The SWLS captures a person's overall cognitive evaluation of their life satisfaction, while the HILS measures a broader sense of psychological balance and integration with one's circumstances. These scales provide structured ground truth targets for the predictive models, both at individual time points and across time.

Linear models have long been shown to be better predictors of psychological constructs than humans [Dawes, 1982, McNemar and Meehl, 1955, Goldberg, 1970]. These models are able to capture simpler underlying relationships between different variables in a relatively straightforward manner, as long as they are linear. Non-linear models, such as artificial neural networks (ANNs), are able to capture complex non-linear relationships between variables, making them potentially even more effective at predicting in the complex realm of psychological constructs and language. The use of state-of-the-art natural language processing (NLP) techniques, such as transformers [Vaswani et al., 2017], has made it feasible to create more accurate and reliable linear and non-linear models for predicting constructs [Kjell et al., 2023a].

Psychological constructs have traditionally been assessed using closed-ended rating scales [Likert, 1932, Robinson, 2014]. These scales have been shown to have limitations, such as a lack of dimensionality and the potential for bias [Fried, 2016]. The use of language-based responses analyzed with artificial intelligence (AI) can complement or even replace these rating scales. Language-based responses contain a lot more data than scalar scale responses [Kjell et al., 2023b]. Combined with the high dimensionality of transformer embeddings and machine learning (ML) prediction models, we might be able to match the dimensionality of constructs with our prediction model and come close to a theoretical upper limit of construct prediction [Kjell et al., 2022b].

The extensive availability of choices for every stage of the prediction procedure presents a considerable challenge in determining the appropriate initial step. This thesis is aiming to explore different preprocessing techniques such as word embedding extraction, principal component analysis (PCA) and scaling, as well as comparing the performance of linear models such as ridge regression and non-linear models like random forests and autokeras re-

gression models through a benchmarking scheme. Comparing these techniques might lead to a better understanding of which approaches are most effective in the realm of construct prediction from word responses, and provide guidance for future studies.

The exploration and development of these prediction models for construct scale score deltas over time, may also lay the groundwork for future improvements on the measurement of construct change and intervention efficiency. With many different measurements of change, or effect size, such as Cohens d, reliable change index (RCI), standardized individual difference (SID), clinical meaningfulness, clinical significance, and statistical significance, and different measurements of effect sizes for different constructs, this work may help lay the foundation in the development of a model that can measure effect sizes in a more uniform way. [Estrada et al., 2019][Lovakov and Agadullina, 2021] [Middel and van Sonderen, 2002]

To improve the accuracy of the prediction process, modern machine learning libraries often incorporate optimization strategies for hyperparameter tuning. Bayesian optimization is a popular approach that has been demonstrated to be highly effective in identifying the optimal set of hyperparameters for a given model. This technique was implemented in this thesis to further refine the prediction models. By reducing the number of hard choices in the prediction process, Bayesian optimization can significantly enhance the efficiency and effectiveness of model performance.

0.1.1 Motivation

The use of transformers in combination with linear and non-linear prediction models has been shown to provide accurate and reliable predictions of construct scale scores, with a correlation coefficient (r) from word responses for HILS (r=0.79) and SWLS (r=0.75) [Kjell et al., 2021b][Kjell et
In addition to their accuracy and reliability, prediction models might have potential to make it easier for patients to interact with the assessment process using natural language, which may be more intuitive and easier to understand than numerical values. Natural language[1] interaction could improve how well patients can express their psychological state and provide more accurate and useful information to researchers.

While score prediction models have been conducted in the past, delta prediction has not been explored in depth. Earlier best practices for score prediction from structured word responses were drawn from neighboring applications that involved prediction from social media data. This thesis aims to fill this gap by conducting both score and delta predictions and exploring the best practices for each, which has not been done in word response scale predictions before.

Depending on geographical and socioeconomic factors, high accuracy psychological evaluations might not be easy to access or be socially stigmatized[Fabrega, 1991]. A digitization of these evaluations might therefore, with its scalability and ability to anonymize, give access to a wider range of potential patients. This digitization could make it possible to give an earlier intervention to many patients and would generate much data for future research, it

[1]Where "natural language" refers to the way humans naturally communicate using words and phrases to convey meaning, as opposed to formal programming languages or mathematical notation.

might also give rise to problems like misinterpretation of results, when no health professional is there to guide the patient and also data privacy and security will be of concern.

0.1.2 Goals

The experiments in this study were designed to further explore the space of psychological construct scale score prediction and investigate the possibility of enhancing prediction accuracy. In addition to score prediction, investigate ways to use NLP, transformers, and regression models to predict score deltas between the same construct evaluation administered at two different times.

A secondary objective of this study is to inform future work on psychological construct prediction by determining the best practices and potential areas for improvement. The comparison of the results from different preprocessing techniques, models, and optimization strategies aims to offer guidance for future studies in this field. To demonstrate the feasibility and accessibility of the presented methods, all predictions in this study are made on a standard consumer-grade computer, table: 8. The ultimate goal is to not only enhance the accuracy of construct predictions, but also to make the prediction process more efficient and accessible for a wider range of potential users without access to high-performance computing resources.

0.1.3 Questions

The study aims to find answers to the following research questions:

1. Can the accuracy of predicting psychological construct scores from natural language data using NLP be improved compared to earlier results and regression methods? And can this be done by switching out the techniques used post data collection?

2. Can NLP and transformers in combination with regression models be used to accurately predict score deltas from text data?

3. What is the impact of different NLP preprocessing techniques, such as the choice of transformer for embedding extraction and PCA, on the prediction accuracy?

4. Which regression models, including ridge regression, Bayesian ridge regression, random forest, autokeras StructuredDataRegressor, and XGBoost, are most effective for word response prediction tasks?

0.1.4 Contributions

This thesis made the following contributions:

- Benchmarked several NLP-based regression pipelines for predicting psychological construct scale scores and score deltas using structured word responses.

- Compared multiple preprocessing strategies, including transformer-based embeddings, PCA, and data scaling.

- Evaluated and compared the performance of several regression models, including ridge regression, Bayesian ridge regression, random forest, XGBoost, and AutoKeras StructuredDataRegressor.

- Demonstrated that AutoKeras consistently outperformed other models across tasks, achieving Pearson r-values above 0.9.

- Introduced and evaluated delta prediction from text, showing it can be accurately modeled using transformer embeddings and engineered features like delta vectors and T1 scores.

- Conducted all experiments on a consumer-grade machine to demonstrate feasibility without server infrastructure.

0.1.5 Related Work

Integration of AI in Psychological Well-being Assessment Efforts to improve the accuracy of psychological well-being assessments have benefited significantly from advances in artificial intelligence, especially through the use of machine learning and natural language processing (NLP) techniques. The work by Kjell et al. (2019) [Kjell et al., 2019a]

stands out, showcasing the ability of AI-based transformers to predict psychological well-being from textual responses with a remarkably high accuracy (Pearson r = .85). This finding demonstrates the potential for machine learning models, especially transformers, to understand contextual nuances in language that may elude traditional survey-based approaches. Complementing this, Lee et al. (2023) [Lee et al., 2023] have broken new ground by leveraging prompt-based generative pre-trained transformers (GPT-3) for automatic item generation in personality tests. These developments not only enhance the precision of psychological assessments but also suggest the potential for AI to contribute significantly in the future of psychological research and practice.

AI and NLP in the Diagnostics of Mental Health Conditions The burgeoning field of NLP has made notable inroads into the diagnostics and monitoring of mental health conditions. A study by [Kjell et al., 2022a] demonstrated the capability of NLP to quantify self-reported symptoms of depression and anxiety through free word responses, reflecting an innovative direction in computational psychiatry. This work is similar to a study by [Eichstaedt et al., 2018], which exploited machine learning to analyze language patterns on social media for depression screening, thereby introducing alternative, data-driven methods for mental health diagnostics. These studies are emblematic of a significant shift towards using AI for the early detection and nuanced understanding of mental health disorders, signaling a future where machine learning tools could potentially augment or even surpass traditional diagnostic methods.

Evaluating the Efficacy of Computational Assessments Over Rating Scales The evaluation of psychological constructs has long relied on traditional closed-ended rating scales; however, recent advances in NLP suggest a reevaluation is due. Research by [Kjell et al., 2019b] provides a compelling argument for the superiority of semantic measures derived from NLP over conventional rating scales, presenting evidence of their higher validity and reliability across various psychological constructs. This perspective gains further support from the findings of [Kjell et al., 2021a], who have demonstrated the enhanced ability of NLP-processed language assessments in capturing the relationship between subjective well-being and cooperative behaviors. Collectively, these studies challenge the status quo and advocate for a nuanced approach that incorporates the rich data available from computational language assessments, potentially leading to a change in how psychological measurements are conducted and interpreted.

Linguistic Indicators of Psychological and Behavioral States The interplay between language and psychology is a central theme in computational psychology, with recent studies reinforcing the idea that linguistic indicators can serve as powerful predictors of psychological and behavioral states. The research by [Kjell et al., 2021a] has been instrumental in correlating language-derived computational assessments with cooperative behaviors, offering a new lens through which subjective well-being can be evaluated. Additionally, the work by [Eichstaedt et al., 2018] underscores the predictive capacity of language used in social media as an early indicator of depression. These pioneering studies advocate for a shift in psychological assessment towards utilizing linguistic data, which may enable more proactive and preventive approaches in mental health care.

0.1.6 Abbreviations

PCA Principal Component Analysis: A statistical method used to reduce the dimensionality of a dataset while retaining as much information as possible. It identifies the directions of maximum variance and transforms the data into a new coordinate system with these directions as the principal components.

SWLS Satisfaction with Life Scale: A psychological rating scale used to measure global satisfaction with one's life. It consists of five statements rated on a 7-point Likert scale, with higher scores indicating greater satisfaction.

HILS Harmony in Life Scale: A rating scale measuring an individual's perception of overall harmony in their life. It consists of 5 items rated on a 7-point Likert scale, emphasizing psychological balance and flexibility.

RCI Reliable Change Index: A standard metric used to evaluate whether the change in a participant's score can be attributed to more than just the error of the measurement tool. It considers the change of a single participant and the reliability of the measurement instrument.

t-SNE t-Distributed Stochastic Neighbor Embedding: A technique used for visualizing high-dimensional data in a lower-dimensional space, typically two or three dimensions. It models the similarity between pairs of data points in the high-dimensional space and then optimizes a low-dimensional embedding that preserves these similarities.

UMAP Uniform Manifold Approximation and Projection: A dimensionality reduction technique that uses an exponential probability distribution in high dimensions and any distance metric, providing a better representation of the data structure without the risk of data leakage.

ANN Artificial Neural Network: Machine learning models that mimic the way the human brain processes information. They consist of interconnected "neurons" that learn to recognize patterns in data using an optimization algorithm called gradient descent.

NLP Natural Language Processing: A field of study focused on the interaction between computers and human languages. It involves processing and analyzing large amounts of natural language data to enable computers to understand and respond to human text and speech.

BERT Bidirectional Encoder Representations from Transformers: A language representation model developed by Google that uses deep bidirectional context to condition on both left and right contexts at all layers. It employs Masked Language Model (MLM) pre-training and Next Sentence Prediction (NSP) to enhance text understanding.

roBERTa Robustly Optimized BERT Pretraining Approach: An extension of BERT, trained on a larger corpus with whole-sentence masking and larger mini-batches. It focuses solely on the MLM strategy without using the NSP loss function.

MLM Masked Language Model: A pre-training objective used in models like BERT, where random tokens in the input sequence are masked, and the model is trained to predict the original identity of these masked tokens.

NSP Next Sentence Prediction: A pre-training task used in BERT to enhance its understanding of text relationships by predicting whether a given sentence follows another sentence in a text.

WWM Whole Word Masking: An advancement in the pre-training of transformer-based language models where entire words are masked instead of subwords, forcing the model to rely solely on contextual information to predict the masked word.

T1/T2 Time points used in longitudinal studies to indicate different measurement occasions, with T1 representing the first measurement and T2 representing the second.

The following section presents the methodology used in this thesis, including preprocessing steps, benchmarking procedures, and model evaluation strategies.

0.2 Theory

This chapter presents the theoretical background and methods used in this thesis. It begins by introducing the psychological constructs and scales used for measurement. Then, it outlines the metrics used to quantify change over time. The chapter concludes with an overview of the preprocessing techniques and machine learning models applied in the experiments.

0.2.1 Constructs

In psychology, a construct is an abstract concept used to describe and understand aspects of human behavior and mental processes that are not directly observable. These constructs, such as intelligence, motivation, or anxiety, are inferred from patterns in behavior, self-reports, or physiological responses. They serve as essential tools for researchers and practitioners to conceptualize and measure complex psychological phenomena.

This paper focuses on two constructs within the domain of well-being psychology: satisfaction with life and harmony in life.

Satisfaction with life The Satisfaction with life scale (SWLS) is a psychological rating scale for measuring global satisfaction with one's life. It consists of five statements that are rated on a 7-point Likert scale, with higher scores indicating greater satisfaction. The scale was developed based on the assumption that satisfaction with life as a whole is a necessary condition for happiness [Diener et al., 1985]. The items included on the scale are listed in Table 1.

List 1: Satisfaction with Life Scale Items

- In most ways my life is close to my ideal.
- The conditions of my life are excellent.
- I am satisfied with my life.
- So far I have gotten the important things I want in life.
- If I could live my life over, I would change almost nothing.

Harmony in life The Harmony in Life Scale (HILS) is a rating scale measuring an individual's perception of overall harmony in their life. It consists of 5 items, each rated on a 7-point Likert-type scale ranging from 1 (strongly disagree) to 7 (strongly agree). The HILS was developed to complement the Satisfaction with Life Scale (SWLS), which focuses on evaluations of actual and expected life circumstances, by emphasizing psychological balance and flexibility [Kjell et al., 2015]. There are variations of the scale, including a shorter version of three items developed later using the items with the strongest predictive ability [Kjell and Diener, 2021]. The items for the five-item scale are presented in List 2.

HILS has been shown to have satisfactory statistical properties and to be correlated with other measures related to well-being [Kjell et al., 2015]. In studies, the HILS has been found to be significantly positively correlated with measures of positive affect ($r = 0.4$, $p \leq 0.001$) and significantly negatively correlated with measures of negative affect ($r = -0.3$, $p \leq 0.01$). In addition, the HILS has been found to explain a larger proportion of the variance in included measures than the SWLS, indicating that it may be a more sensitive tool for assessing subjective well-being [Kjell et al., 2015]. Overall, the HILS is considered a valid and reliable tool for assessing harmony in life.

List 2: Harmony in Life Scale items

- My lifestyle allows me to be in harmony.
- Most aspects of my life are in balance.
- I am in harmony.
- I accept the various conditions of my life.
- I fit in well with my surroundings.

0.2.2 Scales

In psychology, a construct refers to an abstract concept or mental state—such as satisfaction, well-being, or anxiety—that cannot be measured directly but is inferred through self-report or behavioral indicators. Whether evaluating a new treatment, measuring population well-being, diagnosing a patient, or assessing preventive measures, researchers must decide what data to collect and how to analyze it. The standard approach today involves using various self-report scales and interviews. The choice of scale depends on the specific construct being measured, and often multiple scales exist to assess the same construct. [Fried, 2016]

These scales consist of a set of statements rated on a seven-point Likert scale, where participants indicate how much each statement applies to them. The sum of the responses is then calculated to provide an estimate of the construct being measured. If this is done over time with several measurements, the change in these scales is used to estimate population or individual change.

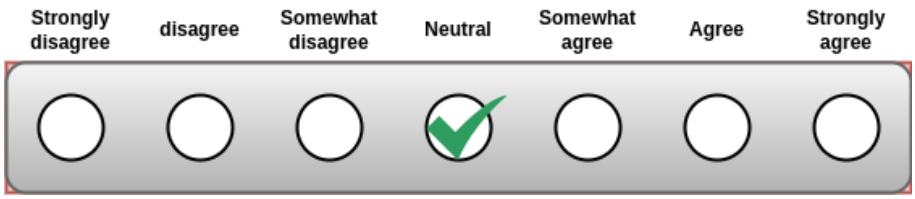

Figure 1: Example of a seven point Likert scale.

The scales provide an estimation of the magnitude of a specific construct within an individual or group. In certain instances, a questionnaire may incorporate multiple scales to measure various constructs simultaneously.

0.2.3 Change metrics

When evaluating the change data from an intervention or survey several metrics can be calculated.

Statistical significance A t-test, with a zero hypothesis that there is no change, produces a p-value that can be used to evaluate if the population change is *statistically significant*. In psychology the change in a data set with a $p \leq 0.05$ is considered *statistically significant*. The t-test takes into account the magnitude of the effect, sample size, effect and instrument reliability and a low p-value minimises the risk of a type one error, finding change when it does not exist. However it does not take into account type two errors, failure to find an existing change.[The Editors of Encyclopedia Britannica, 2022]

Clinical significance Clinical significance refers to a binary classification of change. A participant is considered to have undergone a *clinically significant* change if their pre-intervention score was more than one standard deviation away from the mean of a normal population, and their post-intervention score falls within one standard deviation of that mean [Figure 2].

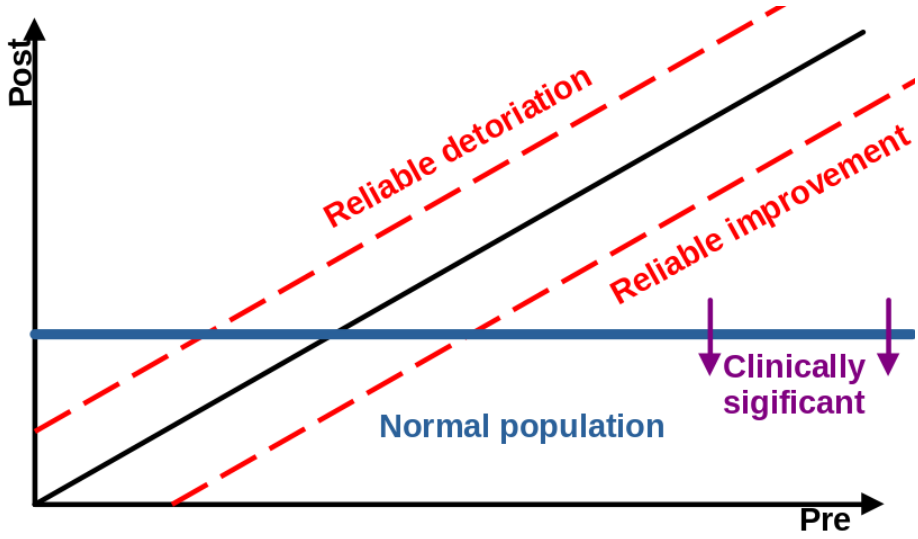

Figure 2: Visualization of change metrics.

Reliable change index The standard metric to evaluate if change can be attributed to more than the error of the measurement tool used is the *reliable change index (RCI)*[Equation 1, Figure 2]. The *RCI* evaluates whether the change observed in a single participant exceeds what could be attributed to measurement error. Reliability of the instrument is estimated using Cronbach's α or the pre-post correlation coefficient, and an RCI score is calculated for each individual data point [Blampied, 2022].

$$RCI = \frac{\Delta_i}{S_{pre} \cdot \sqrt{2(1 - r_{pre-post})}} \geq 1.96 \tag{1}$$

Δ_i = Individual delta
S = standard deviation
$r_{pre-post}$ = pre-post correlation

Cohen's d Cohen's d is a measure of effect size, which is used to indicate the magnitude of the difference between two groups, or the same group at different times, represented by how many standard deviations the mean of the sample has changed. It is calculated by dividing the difference between the means of the two groups by the pooled standard deviation. This can be expressed mathematically as:

$$d = \frac{\bar{x_1} - \bar{x_2}}{s_p}$$

where $\bar{x_1}$ and $\bar{x_2}$ are the means of the two groups, and s_p is the pooled standard deviation, which is calculated as:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

where n_1 and n_2 are the sample sizes of the two groups, and s_1 and s_2 are the standard deviations of the two groups. This measure allows researchers to easily compare the magnitude of the difference between two groups, even if they have different sample sizes or variances [Estrada et al., 2019].

Pearson's correlation coefficient (r) Pearson's r is a measure of the linear correlation between two variables. It ranges from -1 (perfect negative linear correlation) to +1 (perfect positive linear correlation), with 0 indicating no linear correlation. In this study, it is used as the primary evaluation metric to assess how well the predicted construct scores align with the actual scores.

0.2.4 Data sets with change metrics

The data set was collected, through a website called Prolific, from a normal population with no applied intervention, this can be seen in the figures 3,4 and 5 where the delta distribution centers around zero and very few participants in the study has made a change that is reliable according to RCI. Data was collected two times for each participant with a fixed time in between collections for the delta value.

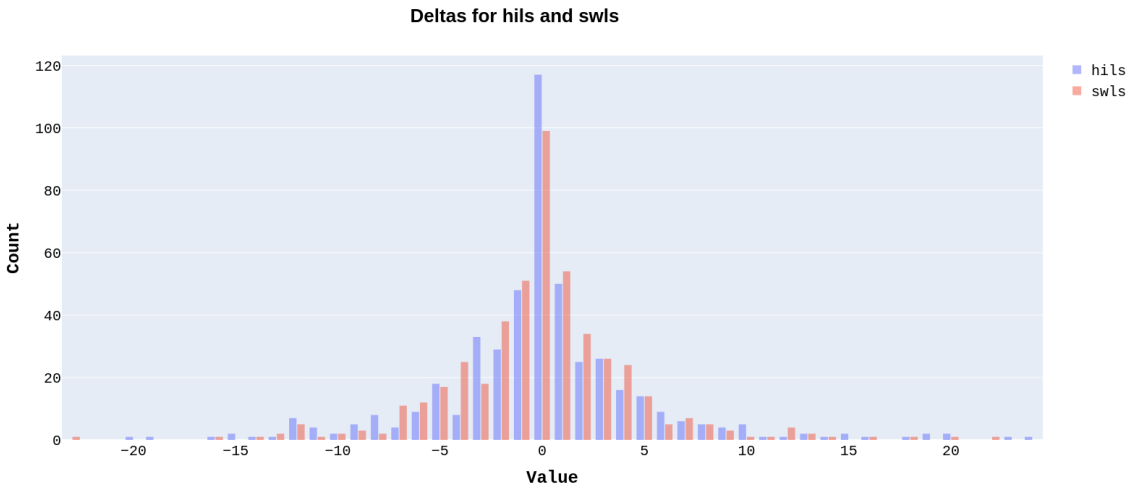

Figure 3: Histogram of the t1, t2 scale score deltas.

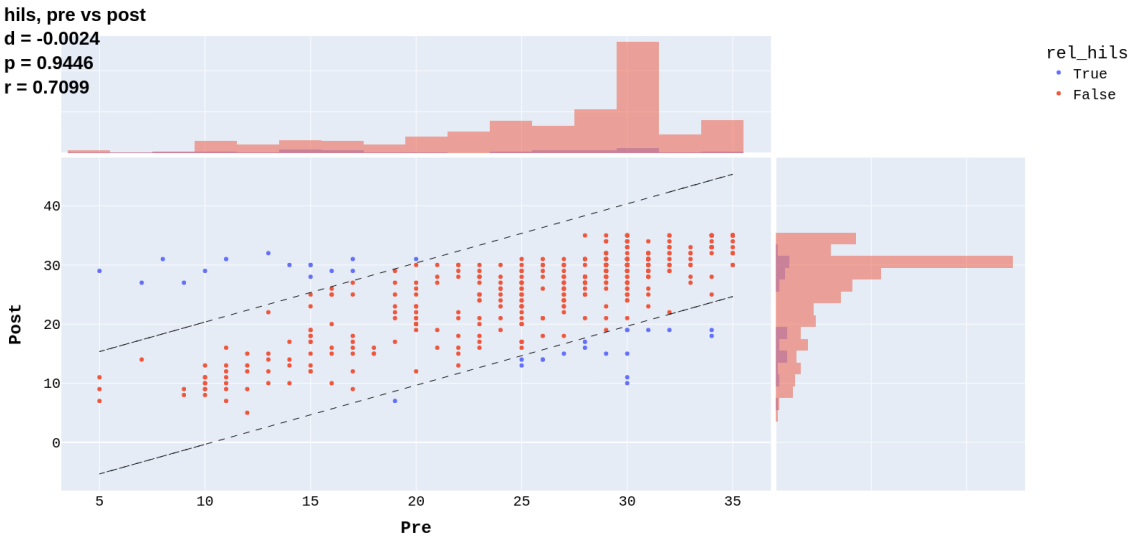

Figure 4: Pre-post scatter plot for HILS with change metrics.The rel_hils tels you if the data point has gone through reliable change according to RCI.

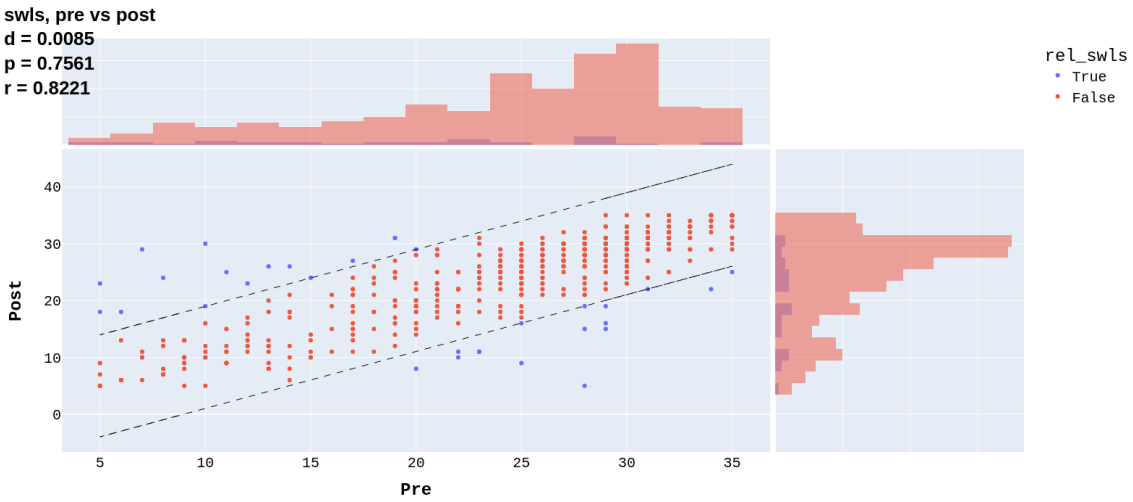

Figure 5: Pre-post scatter plot for SWLS with change metrics. The rel_swls tels you if the data point has gone through reliable change according to RCI.

0.2.5 Preprocessing techniques

Principal component analysis (PCA) Principal Component Analysis (PCA) is a statistical method used to identify patterns in data. It involves reducing the dimensionality of a data set while retaining as much information as possible. The process starts by identifying the directions of maximum variance in the data and transforming the data set into a new coordinate system with these directions as the principal components. The first component, also known as the first principal component, captures the most variance, and

each subsequent component captures successively less variance. The transformed data, with its reduced dimensionality, may not represent the data with perfect accuracy, but it requires fewer calculations, leading to simpler models for predictions. When the dimensionality is low enough, the data can be visualized in this lower-dimensional space, making it easier to identify patterns and correlations.

Embedding extraction The embeddings are vectors of floats that capture the semantic information of the input text data. Each layer of the transformer generates a float vector embedding representation of the text. In this study, as well as in related work, the embeddings from the last two layers of selected transformer models are used. These embeddings are used as features for various NLP tasks, including score and delta prediction. The last two layers of the models were selected because they contain rich semantic information about the input text data. The advantage of using embeddings as features is that it is possible to do vector operations on them, allowing the models to gain a better understanding of the natural language input and make more accurate predictions.

0.2.6 Machine Learning Models

Ridge Regression

Ridge regression is a linear modeling technique that extends ordinary least squares (OLS) regression by adding an L2 regularization term. Ordinary Least Squares is a method for estimating the coefficients of a linear regression model by minimizing the sum of squared differences between the observed values and the values predicted by the model. L2 regularization adds a penalty based on the square of the magnitude of the coefficients, encouraging smaller weights and reducing the risk of overfitting. This regularization helps prevent overfitting, when a model captures noise rather than true patterns, especially in situations where predictor variables are highly correlated or when the number of predictors exceeds the number of observations. The objective function for ridge regression is:

$$J(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2 + \lambda\|\boldsymbol{w}\|_2^2 \tag{2}$$

Here:

- $\boldsymbol{w}$ is the vector of model coefficients.
- $\boldsymbol{x}_i$ is the feature vector for the $\boldsymbol{i}$-th observation.
- y_i is the target value for the $\boldsymbol{i}$-th observation.
- n is the number of observations.
- $\lambda \geq 0$ is the regularization parameter controlling the strength of the penalty.

The first term measures the fit of the model using mean squared error (MSE), which is the average of squared differences between predicted and true values. The second term penalizes large coefficients, effectively shrinking them towards zero to reduce complexity.

Bayesian Ridge Regression

Bayesian Ridge Regression extends ridge regression by estimating a probability distribution over the model parameters instead of point estimates. This provides a measure of uncertainty in the predictions. A Gaussian prior is placed over the coefficients:

$$p(\boldsymbol{w}) = \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{I}) \tag{3}$$

and the likelihood of the observed data is also assumed to be Gaussian:

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{w}) = \mathcal{N}(\mathbf{X}\boldsymbol{w}, \alpha^{-1}\mathbf{I}) \tag{4}$$

Bayes' theorem combines prior and likelihood into a posterior distribution:

$$p(\boldsymbol{w} \mid \mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{w})\, p(\boldsymbol{w})}{p(\mathbf{y} \mid \mathbf{X})} \tag{5}$$

The result is a Gaussian posterior over the weights. This probabilistic framework makes the model robust to overfitting, especially in small datasets, and provides credible intervals around predictions.

Random Forest

Random Forest Regression is an ensemble method that builds multiple decision trees and averages their outputs. It helps reduce overfitting by lowering model variance. Each tree is trained on a bootstrapped sample, a randomly selected subset of the data with replacement. At each node, a random subset of features is used to determine the best split, reducing correlation between trees. The training steps are:

1. Generate several bootstrapped datasets.

2. Train one decision tree on each:

    - Select random features at each node.
    - Split based on criteria such as mean squared error.

3. Average the predictions from all trees:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(\boldsymbol{x}) \tag{6}$$

Where $\boldsymbol{T}$ is the number of trees and $\boldsymbol{h_t(x)}$ is the $\boldsymbol{t}$-th tree's prediction. Random Forests can also output feature importances and handle high-dimensional data well.

XGBoost

XGBoost (Extreme Gradient Boosting) builds decision trees sequentially, where each tree tries to correct the errors of the previous ones. It is known for high accuracy, speed, and built-in regularization. The objective function includes both a loss term and a regularization term to penalize model complexity:

$$\mathcal{L} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{7}$$

where:

- L is the loss function (e.g., squared error).
- $\Omega(f_k)$ is the regularization term defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{8}$$

- T is the number of leaves, w_j is the score on each leaf, γ and λ are regularization parameters.

To optimize this, XGBoost uses a second-order Taylor expansion of the loss function, allowing more accurate estimation of gradients and efficient learning. L1 regularization can set weights to zero (feature selection), while L2 shrinks them (as in ridge).

ANN's

Artificial neural networks Artificial neural networks are machine learning models that mimic the way the human brain processes information. They consist of interconnected "neurons" that can learn to recognize patterns in data. The conventional way to train a neural network is to use an optimization algorithm called gradient descent. This algorithm adjusts the signal strength, weight, between neurons in the network to minimize a cost function, which measures the error between the output of the neural network and the true values we are trying to predict. The weights are updated according to the gradient of the cost function, which tells us the direction in which to adjust the weights to reduce the error. This process is repeated for a number of iterations, called epochs or batches, until the cost function reaches a minimum and the neural network has hopefully learned to make accurate predictions. In code, the implementation of gradient descent involves defining the cost function and its gradient, initializing the weights of the network, and then iteratively updating the weights according to the gradient descent formula in a loop for a specified number of epochs.

$$w_i = w_i - \alpha \frac{\partial J}{\partial w_i} \tag{9}$$

Transformers

Transformers are a neural network architecture introduced by Vaswani et al. in "Attention is All You Need" [Vaswani et al., 2017], designed to handle sequence data using attention mechanisms rather than recurrence. This allows for efficient parallelization and better handling of long-range dependencies, making transformers particularly effective in natural language processing (NLP).

Architecture Overview The architecture consists of two main components: an encoder that processes the input sequence into a contextual representation, and a decoder that generates the output sequence using this representation. Unlike previous models, transformers process all tokens in parallel rather than sequentially, significantly improving scalability.

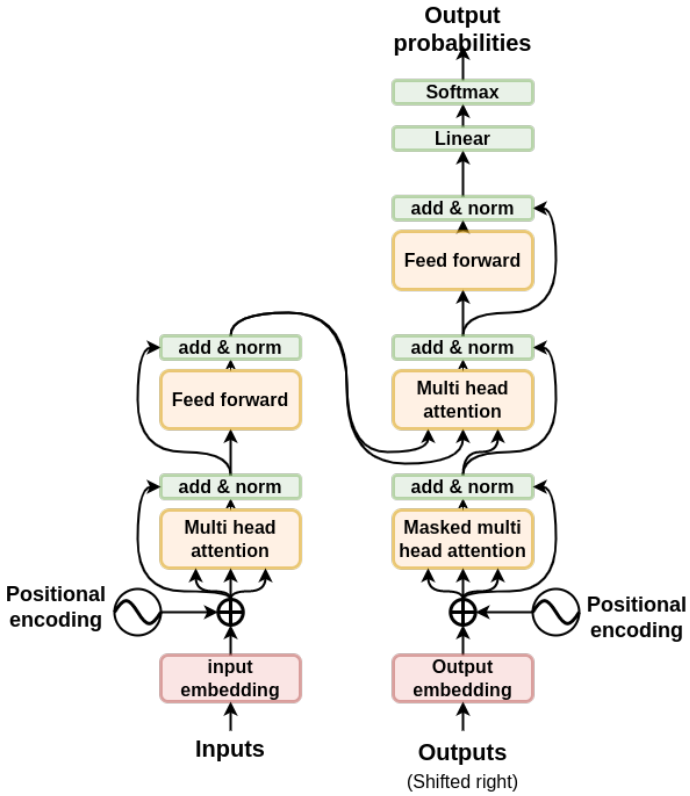

Figure 6: Transformer encoder-decoder structure.

Attention Mechanism The core innovation is the self-attention mechanism, which allows the model to assign different weights to tokens in the sequence based on their relevance to one another. For each position, queries (Q), keys (K), and values (V) are computed, and attention is calculated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{10}$$

This enables the model to contextualize each word in relation to the entire input.

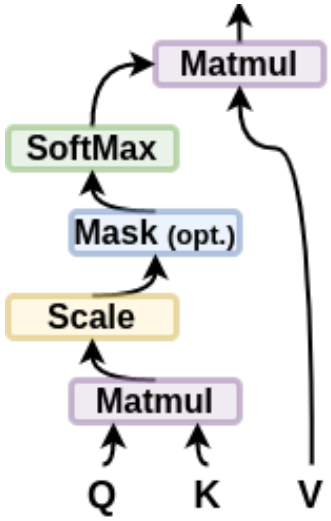

Figure 7: Self-attention mechanism in the transformer.

Multi-Head Attention To enrich this process, transformers use multi-head attention, where multiple sets of attention mechanisms run in parallel, each focusing on different aspects of the sequence. The outputs from all heads are concatenated and linearly transformed:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{11}$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{12}$$

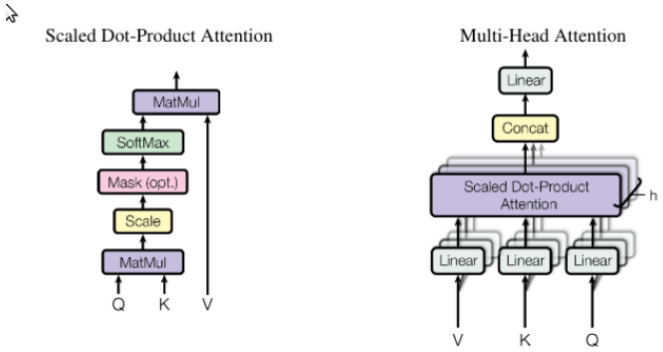

Figure 8: Multi-head attention mechanism. Each head learns to focus on different parts of the input sequence.

Positional Encoding Since transformers do not inherently model word order, positional encodings are added to input embeddings to introduce sequence information. These are computed using sinusoidal functions:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right), \quad PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{13}$$

Impact Transformers have become foundational in modern NLP systems, enabling models such as BERT, GPT, and T5. Their influence now extends to other domains, including vision and bioinformatics. The encoder-decoder structure and attention mechanisms are visually summarized in Figures 6 and 7.

BERT BERT (Bidirectional Encoder Representations from Transformers) [Devlin et al., 2018] is a transformative language representation model developed by Google. It diverges from traditional unidirectional context models by utilizing deep bidirectional context, conditioning simultaneously on both left and right contexts at all layers. Derived from the Transformer model [Vaswani et al., 2017], BERT is distinguished by its Masked Language Model (MLM) pre-training objective, where it predicts the original identity of randomly masked input tokens, and its Next Sentence Prediction (NSP), which enhances its understanding of text relationships. BERT's pre-trained model, adaptable with minimal task-specific modifications, establishes state-of-the-art performance across numerous NLP tasks.

roBERTa RoBERTa, a Robustly Optimized BERT Pretraining Approach, extends BERT's transformer-based methodology for NLP tasks. It differentiates itself by training on a vastly larger corpus of over 160 billion words and by employing "whole word masking" for a more effective training process [Liu et al., 2019].

Differences in training between BERT and roBERTa:

(a) RoBERTa applies whole-sentence masking and employs larger mini-batches compared to BERT's token-level masking and smaller batches.

(b) It is trained on a more extensive dataset with a broader byte-level BPE vocabulary.

(c) Unlike BERT, roBERTa does not utilize the NSP loss function, instead focusing solely on the MLM strategy.

Masking Whole word masking (WWM) is an advancement in the pre-training of transformer-based language models, particularly enhancing the masked language modeling (MLM) technique introduced with BERT. Traditional MLM approaches might mask subwords or pieces of a word, leading to situations where the model still has partial clues to predict the masked token. WWM tackles this issue by masking all subtokens of a chosen word simultaneously, compelling the model to rely solely on contextual information to predict the full masked word. For instance, instead of masking just "play" or "##ing" in the word "playing", WWM masks the entire word as "[MASK] [MASK]", pushing the model to understand context at a whole-word level rather than predicting parts of a word in isolation. This methodology is particularly beneficial for the training of models like roBERTa, which has demonstrated improved performance across a spectrum of NLP tasks due to this more rigorous and context-dependent pre-training strategy.

0.2.7 Cross-validation

Cross-validation is a resampling method used to assess the generalization performance of machine learning models, particularly when working with limited data. In k-fold cross-validation, the dataset is randomly partitioned into k equally sized folds. The model is trained

k times, each time using *k-1* folds for training and the remaining fold for validation. This ensures that every data point is used for both training and evaluation exactly once.

The final performance metric is computed as the average over all *k* runs, which provides a more reliable estimate of how the model is expected to perform on unseen data compared to a single train-test split. Cross-validation reduces the variance associated with data partitioning and helps mitigate the risk of overfitting, especially in high-variance models or small datasets.

0.2.8 Bayesian Hyperparameter Optimization

Bayesian optimization is an efficient method for tuning hyperparameters by building a probabilistic model of the function mapping hyperparameters to performance. It selects new configurations using an acquisition function that balances exploration of the search space and exploitation of known good regions. This allows for faster convergence to optimal model settings compared to random or grid search.

0.3 Method

This chapter describes the methodology used to predict psychological construct scores and their changes over time based on participants' word responses. The goal is to explore whether short, self-generated textual descriptions can be used to infer scores from validated psychological scales. The input to all models is text data (ten words per construct, per time point) and the output is either a predicted scale score (at a specific time point) or a predicted change score (delta between T1 and T2). This chapter outlines the dataset used, preprocessing steps, regression models, and the benchmarking procedure for evaluation.

0.3.1 Dataset

HILS and SWLS t1 and t2 dataset used in this study consisted of 477 participants and included harmony and satisfaction word data and HILS (Harmony in Life Scale) and SWLS (Satisfaction with Life Scale) scores for both time 1 (T1) and time 2 (T2). This is the same dataset used in Kjell et al. (2019), with new model evaluations performed in this thesis. The data consists of a 10-word description of the psychological construct paired with a scale score for the same construct, collected two times using the online platform Prolific, with 30.79 days, (SD=2.01), between t1 and t2[Kjell et al., 2019a]. The word responses were collected prior to the rating scales, so the rating scales could not influence the word responses but the other way was possible.

_hilstotalt1	_hilstotalt2	harmony_t1	harmony_t2
32.0	19.0	synchronized collaborative focused determ...	somewhat loved anxious amazed hopeful tr...
31.0	32.0	happy optimistic fulfilled complete ne...	happy energetic creative thoughtful ac...
20.0	25.0	balance chaos order bills debt insomn...	happy balanced angelic centred peacefu...
32.0	34.0	Peace Love Truth Justice Freedom Law Na...	Peace Love Truth Justice Freedom Univer...
15.0	18.0	tired content love taco distressed fi...	content calm cogent rested thoughtful in...

_swlstotalt1	_swlstotalt2	satisfaction_t1	satisfaction_t2
29.0	16.0	Happy content determined proud boastful ...	failure loser waste potential anxious wa...
28.0	30.0	happy satisfied fulfilled complete ho...	happy energetic peaceful serene though...
16.0	15.0	free fun intelligent riveting refreshing...	regret anxiety remorse debt crisis apartm...
30.0	33.0	Truth Love Peace Justice Freedom Power ...	Happy Kindhearted Nurturing Determined Am...
15.0	11.0	content worried masked debt candour okay ...	frustrated confused upset annoyed insult...

Figure 9: The structure of the data set, showing five data points, where the first two columns is the total scale scores for T1 and T2 respectively and the last two columns is the ten word description of the construct provided by the participants at T1 and T2 respectively as well. The top table shows anonymized data for the Harmony In Life construct and the bottom one for the satisfaction with life construct.

0.3.2 Data analysis

For the data analysis, histograms and scatter plots of the score and delta data was produced with the python libraries Pandas and Plotly. These plots were produced to see how the data is distributed and to look for outliers and trends in the data. Also the change metrics RCI and Cohen's d is applied to the data set and visualised. The visualization of RCI can help analyzing how many individual data-points that has made a statistically significant change between T1 and T2, and Cohen's d to measure the same on a group level.

0.3.3 Data preprocessing

In the preprocessing stage of the experiment the embeddings that was used as the input features is extracted from four different transformer models. The embeddings are extracted from the hidden state of the last two layers of the transformer model and PCA for all the dimensions in figure 11 is applied. The process involved the computation of the mean value by summing up the embeddings of each token in the layer and calculating their mean coordinates in embedding space. These individual token-level means were then concatenated together and saved to disk.

For delta prediction a delta vector dataset is also prepared, where the delta vector for each of the extracted layers is calculated with vector subtraction,

$$T2_{emb-lay} - T1_{emb-lay} = Delta_{lay}$$

A more extensive list of the preprocessing done can be seen in table 3

1. Removing unnecessary columns: Any columns that were not relevant to the analysis were removed from the dataset. Only the columns with T1 and T2 construct score and word response where used for analysis.

2. Text embeddings: BERT and RoBERTa large and base models were used to extract embeddings from the harmony and satisfaction word data in the dataset. The last two layers of the models were concatenated. The length of these embedding vectors is 768 for the base models and 1024 for the large models.

3. PCA of embeddings: Principal Component Analysis (PCA) was applied to the text embeddings to reduce the dimensionality of the data. The last two layers were concatenated after PCA.

4. Additional feature options: In order to predict deltas, a delta vector option used to represent the movement between T1 and T2 was calculated as the vector between the T1 embedding vector and the T2 embedding vector in the embedding space. Additionally T1 scores could be added as the first feature in the delta vector.

Table 3: The preprocessig done to the data set.

0.3.4 Benchmarking

Models and parameters

To evaluate the predictive performance, a series of controlled benchmarking experiments were conducted. Each experiment tested a specific combination of:

- **Transformer model:** BERT-base, BERT-large, RoBERTa-base, RoBERTa-large
- **Prediction target:** Scale score at T1/T2 (score prediction), or difference between T1 and T2 (delta prediction)
- **Construct:** Harmony in Life Scale (HILS) or Satisfaction with Life Scale (SWLS)
- **Model type:** Ridge regression, Bayesian ridge regression, Random Forest, XGBoost, or feedforward ANN
- **PCA dimensionality:** [8, 16, 32, 64, 128] (varied per experiment)
- **Delta vector option:** Whether embeddings were concatenated or subtracted (for delta tasks)
- **T1 score input:** Whether the T1 score was included as a numerical feature in delta prediction

For each unique configuration, embeddings were extracted and reduced using PCA, followed by 5-fold cross-validation. Model-specific hyperparameters were optimized using Bayesian optimization (described below). Evaluation metrics (Pearson's r, MAE, MSE, RMSE) were recorded for each fold and averaged.

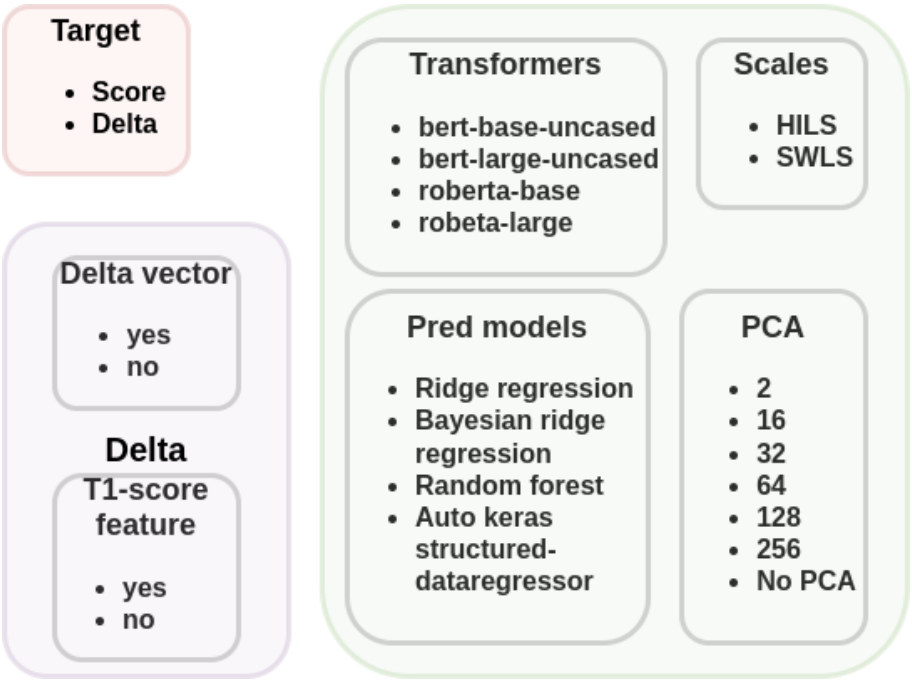

Figure 10: List of Parameters in the Benchmark Scheme where the parameters in the green box is for both delta and score prediction while the parameters in the purple box is only used in delta prediction.

In the benchmarking scheme, the process of optimizing the hyperparameters of the regression models is a crucial step. To do this step bayesian optimization strategies was implemented through the existing python libraries, Sklearn and autokeras, to optimize the hyperparameters for each model.

Model naming convention

In the result graphs and saved files, model configurations are represented using short codes. For example:

- `ak_struct_reg`: AutoKeras structured data regression model
- `xgb`: XGBoost
- `rf`: Random Forest
- `bayridge`: Bayesian Ridge Regression
- `ridge`: Ridge Regression

These codes are combined with transformer and task-specific identifiers to form full labels (e.g., `roberta_large_hils_delta_ak_struct_reg`).

Evaluation procedure

The performance of the different models is compared using a 5-fold cross-validation scheme. The mean result across folds is saved for further analysis. The benchmark is performed for all combinations of the target construct, transformer model, rating scale, regression model, PCA dimensionality, inclusion of the delta vector, and the characteristic of the T1 score, as shown in Figure 10. The models are evaluated primarily using Pearson's correlation coefficient (r), along with standard metrics such as mean absolute error (MAE), mean squared error (MSE) and root mean squared error (RMSE).

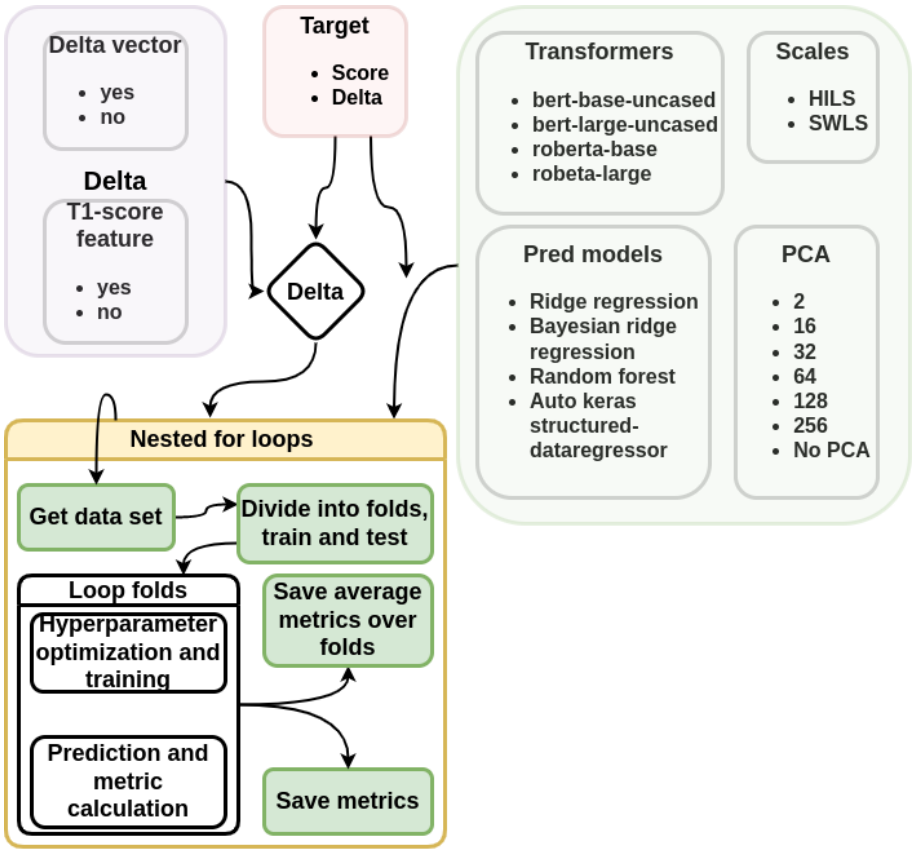

Figure 11: A diagram of the benchmarking procedure.

A diagram of the benchmarking process can be seen in Figure 11 and the code can be found in the github repository..

Score prediction

For the score prediction benchmark, a cross-validation training sequence with Bayesian hyperparameter optimization was performed for each unique parameter combination in Table 10. PCA and embedding extraction and concatenation was done prior to training, this means the PCA model was trained on the entire data set. The evaluation metrics was saved for each fold and the average over folds calculated as the final results used in this thesis. The individual fold results was also saved and can be found in the git repository.

Delta prediction

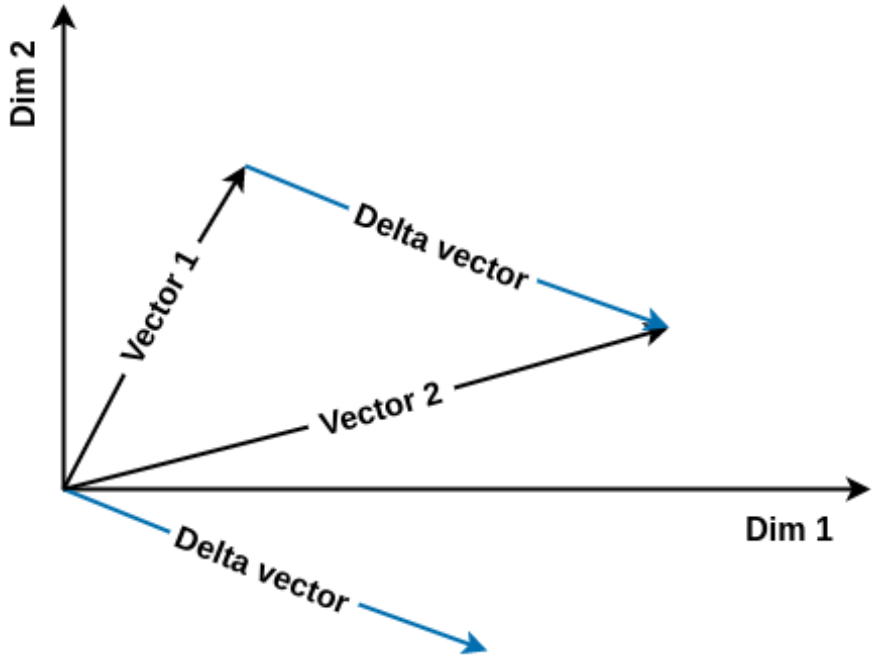

Figure 12: Visualisation of the delta vector in only two dimensional embedding space.

For the delta prediction task, the same benchmarking scheme as in the score prediction task was used, with the parameters outlined in figure 10. However, two additional options were included in the pipeline: the delta vector and the T1 scale score as an input feature. By default, the models were trained on the concatenated embeddings of T1 and T2 words. With the delta vector option, the input for the models is the delta embedding vector, which is the vector between the embeddings for T1 and T2 in embedding space (as shown in equation 14 and figure 12). The added T1 feature option adds the T1 scale score value as the first input feature.

$$\vec{\delta}_{Embs} = \vec{t2}_{Embs} - \vec{t1}_{Embs} \tag{14}$$

0.3.5 LLM-Enhanced Authoring Workflow

To support the writing process of this thesis, I employed large language models (LLMs) as interactive tools throughout multiple stages. Using retrieval-augmented generation (RAG), I indexed all my notes and draft content in a vector database, allowing me to query and contextualize relevant material efficiently. I iteratively engaged with a variety of LLMs (e.g., GPT-3.5, GPT-4, GPT-4o, GPT-4o-mini, Gemini 1.0–2.5, Claude, and others), crafting system prompts to critically examine arguments, improve clarity, and refine phrasing. These models were used both for brainstorming and for sanity-checking existing content. I continued this back-and-forth until the refinement level made manual editing more effective than further prompting. After manual revisions, I often returned to the LLMs for final reviews,

comparing suggestions across models to identify potential improvements and linguistic alternatives.

0.4 Results

This section presents the results of the predictive modeling benchmarks. Both score and delta predictions were evaluated using a range of transformer-based embeddings and regression models. The performance is assessed using several metrics, and key findings are interpreted in relation to the research questions.

0.4.1 Score prediction

The results of the score prediction experiment presented in the following graphs are based on the combination of techniques utilized. Both t1 and t2 values were used to increase the dataset size, and the PCA was performed on each layer of the transformer model embeddings. The last two layers of each transformer model were utilized for prediction and Bayesian hyperparameter optimization was applied to each prediction model. The models were trained using a 5-fold cross validation scheme, with the average result serving as the outcome for each metric. The top 10 combinations for each scale are also presented in a table.

In figures 13 and 14 the x-axis represents the regression models used or the transformer from which the embeddings were extracted. The y-axis represents the Pearson's correlation coefficient (r-value) between the predicted and true scores. The plots are colored based on whether PCA was used in preprocessing or not. There are two types of plots for each scenario: a scatter plot and a box plot. The scatter plot shows the distribution of the r-values for each model or transformer, while the box plot shows the distribution of the r-values along with their median and quartiles. These plots provide a visual comparison of the performance of different regression models and transformers in predicting the psychological scale scores.

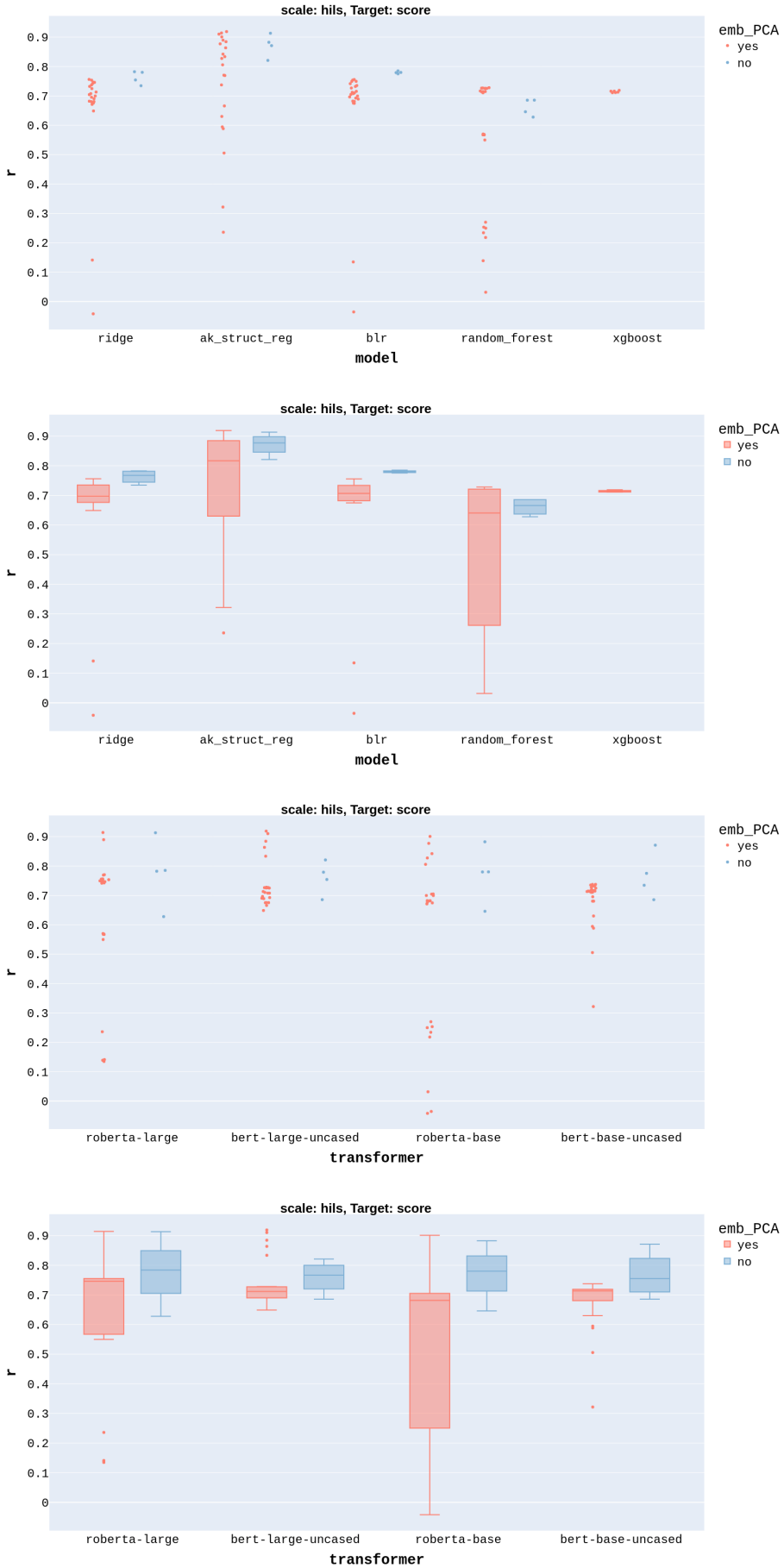

Figure 13: HILS score prediction results, colored by embedding PCA application. The top two plots show performance by regression model; the bottom two show performance by transformer.

model	mae	mse	rmse	r2	r	elapsed_time	n_features	transformer	PCA	pca_n
ak_struct_reg	0.996	7.123	1.302	0.845	0.929	707.187	64	bert-base-uncased	yes	32
ak_struct_reg	1.598	6.812	2.251	0.854	0.923	1800.454	512	roberta-large	yes	256
ak_struct_reg	1.392	7.102	1.755	0.846	0.921	1325.291	256	bert-base-uncased	yes	128
ak_struct_reg	1.559	6.732	2.038	0.855	0.919	1066.675	128	bert-large-uncased	yes	64
ak_struct_reg	1.491	6.824	2.047	0.853	0.919	2030.117	512	bert-large-uncased	yes	256
ak_struct_reg	1.864	8.195	2.685	0.826	0.914	1250.641	256	roberta-large	yes	128
ak_struct_reg	2.053	8.589	2.804	0.818	0.913	6333.132	2048	roberta-large	no	2
ak_struct_reg	1.692	7.682	2.357	0.835	0.910	2647.678	512	bert-large-uncased	yes	256
ak_struct_reg	1.511	8.054	2.028	0.826	0.901	1970.726	512	roberta-base	yes	256
ak_struct_reg	2.477	11.353	3.288	0.761	0.890	2270.643	512	roberta-large	yes	256

Table 4: Top ten results for HILS score prediction sorted on Pearson's r-value. where ak_struct_reg is the models created with the AutoKeras StructuredDataRegressor package

In addition to Pearson's r, other metrics such as MAE (mean absolute error), MSE (mean squared error), RMSE (root mean squared error), and R^2 (coefficient of determination) were included. While r shows correlation strength, RMSE gives the average prediction error, and R^2 indicates how much variance in the true values is explained by the model.

Interpretation of HILS Score Prediction

Interpretation – HILS score prediction Among the transformer models, bert-base-uncased and roberta-large achieved the highest r-values in HILS score prediction, with PCA generally improving performance. The AutoKeras StructuredDataRegressor (ak_struct_reg) model clearly outperformed all other regression models, consistently achieving r-values above 0.90. In contrast, traditional models like Ridge and Random Forest performed worse and showed greater variance. This indicates that neural network-based models are better at capturing the underlying structure of the word embeddings.

The graphs illustrate Pearson's correlation coefficients (r) for different prediction models and transformers used to generate embeddings. The lower graph compares transformer models such as roberta-large, bert-large-uncased, roberta-base, and bert-base-uncased. The points are color-coded, with red indicating the application of Principal Component Analysis (PCA) to the embeddings (emb_PCA: yes) and blue for embeddings without PCA (emb_PCA: no). The pattern suggests that PCA application does not consistently affect the Pearson's correlation coefficient across different transformer models.

In contrast, the upper graph contrasts various prediction models: ridge, ak_struct_reg, blr, random_forest, and xgboost, using the same color scheme for PCA application. Here, the ridge and ak_struct_reg models with PCA (red points) appear to achieve higher Pearson's correlation coefficients, whereas models like blr, random_forest, and xgboost show a wider spread of coefficients regardless of PCA application. The table lists the top ten HILS score prediction results, where each entry corresponds to a model developed with the AutoKeras StructuredDataRegressor package, denoted as 'ak_struct_reg'. The models are ranked by their Pearson's correlation coefficient (r-value), with additional performance metrics including mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared (r2) value.

Each model's performance is reported alongside the elapsed time for computation, the

number of features used (n_features), the transformer model applied for embeddings, and whether Principal Component Analysis (PCA) was employed. The number of PCA components is also provided (pca_n).

Models span a range of configurations, using transformers like ‘bert-base-uncased‘, ‘bert-large-uncased‘, ‘roberta-large‘, and ‘roberta-base‘, with varying numbers of features and PCA components. The majority of the models have applied PCA to the embeddings, with the exception of one entry, where PCA was not utilized.

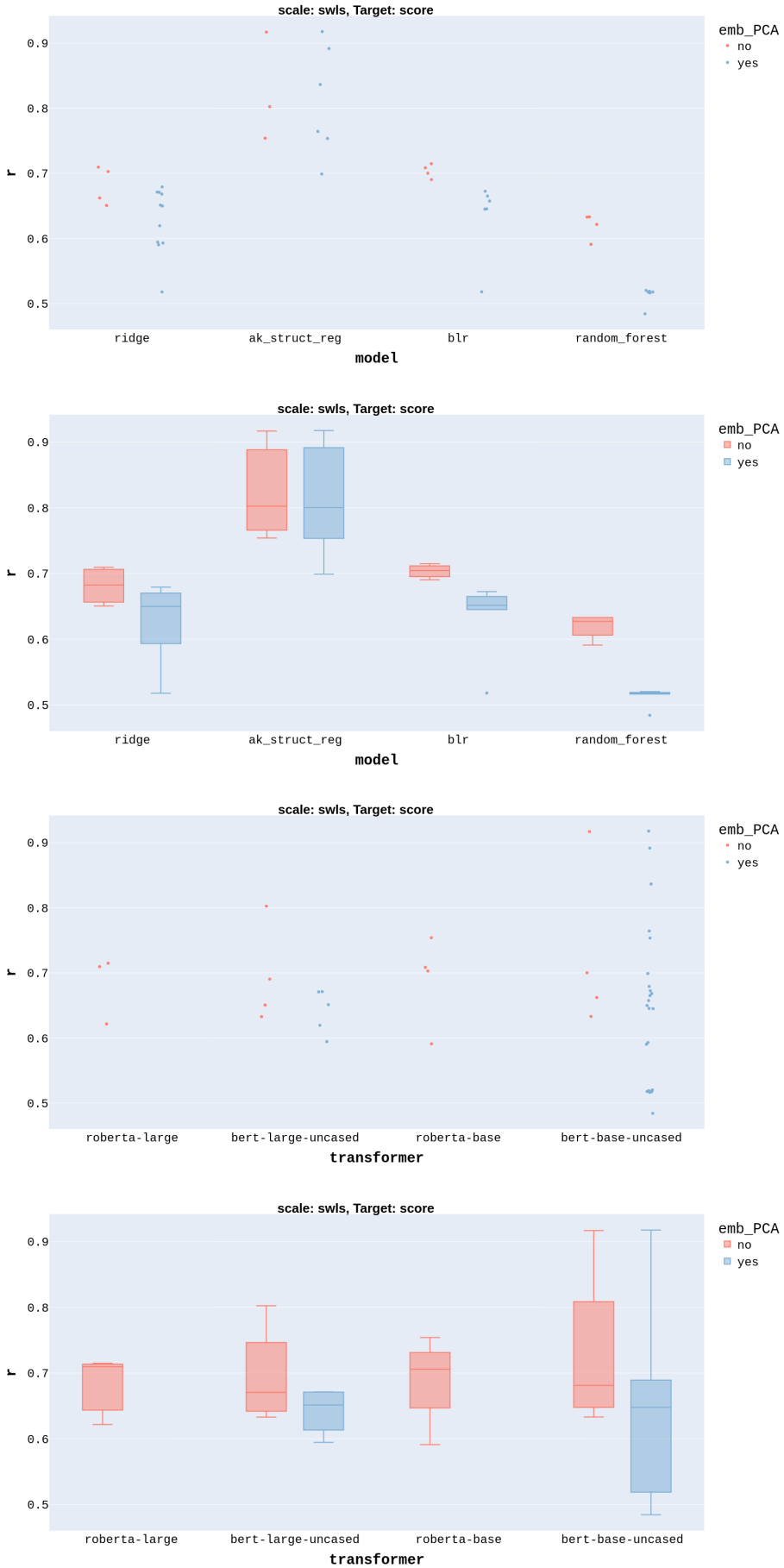

Figure 14: SWLS score prediction results, colored by embedding PCA application. The top two plots show performance by regression model; the bottom two show performance by transformer.

Among transformer models, bert-base-uncased and roberta-large consistently achieved higher r-values across both scales. As for regression models, AutoKeras StructuredDataRegressor clearly outperformed traditional models such as Ridge and Random Forest, suggesting that deep learning approaches are better suited for this task.

model	mae	mse	rmse	r2	r	elapsed_time	n_features	transformer	PCA	pca_n
ak_struct_reg	1.674	8.858	2.219	0.857	0.918	2210.682	512	bert-base-uncased	yes	256
ak_struct_reg	2.328	10.414	2.974	0.827	0.917	4953.432	1536	bert-base-uncased	no	2
ak_struct_reg	2.526	13.105	3.295	0.784	0.892	1217.120	256	bert-base-uncased	yes	128
ak_struct_reg	3.329	18.820	4.224	0.685	0.837	693.073	64	bert-base-uncased	yes	32
ak_struct_reg	3.573	21.153	4.557	0.642	0.802	8056.211	2048	bert-large-uncased	no	2
ak_struct_reg	4.132	25.781	5.044	0.563	0.764	961.044	128	bert-base-uncased	yes	64
ak_struct_reg	3.939	25.508	5.045	0.567	0.754	5673.468	1536	roberta-base	no	2
ak_struct_reg	3.917	25.817	5.029	0.566	0.754	467.260	4	bert-base-uncased	yes	2
blr	4.285	29.087	5.385	0.506	0.715	124.926	2048	roberta-large	no	2
ridge	4.299	29.542	5.425	0.499	0.710	57.514	2048	roberta-large	no	2

Table 5: Top ten results for swls score prediction sorted on Pearson's r-value. where ak_struct_reg is the models created with the AutoKeras StructuredDataRegressor package

The bert-base-uncased transformer achieved the best results for SWLS score prediction, with several configurations exceeding an r-value of 0.91. Unlike HILS, PCA had a more mixed effect on model performance, and some non-PCA models achieved near-top scores. Although $ak_struct_reg models consistently topped the list, traditional models like Ridge and BLR appeared in the top 10 values and higher errors.$

Interpretation of SWLS Score Prediction

Interpretation – SWLS score prediction These results support RQ1 by showing that prediction accuracy for HILS scores can be significantly improved using transformer embeddings and AutoKeras models. The strong performance of models incorporating T1 scores and PCA also supports RQ2, confirming that delta prediction using text data is feasible and accurate. The findings further show that preprocessing strategies, such as PCA and delta vectors, have measurable impact, addressing RQ3, while the superior performance of AutoKeras over Ridge, Random Forest, and XGBoost confirms RQ4.

0.4.2 Delta prediction

The results of the delta prediction experiment were obtained using the same dataset as the score prediction, with the delta between t1 and t2 being the target for prediction. The same testing scheme was employed, with the added consideration of using either the concatenated t1 and t2 embeddings or a delta vector (t2 embeddings - t1 embeddings) as input for the models, both with and without the t1 value included in the input.

In the figures 15 and 16, the delta prediction is presented through a grid of six plots. The top two plots show a scatter plot and a box plot, respectively, with Pearson's correlation coefficient (r) on the y-axis and regression models on the x-axis. The plots are colored based

on the application of PCA (yes/no). These plots provide a visual representation of the correlation between the predicted and observed delta scores, with higher values of r indicating a stronger correlation. The box plots complement the scatter plots by showing the distribution of the data within each group.

The middle two plots are also scatter and box plots, with r on the y-axis and transformers used for embedding extraction on the x-axis. These plots are colored based on the application of the added T1 score feature (yes/no). These plots help in understanding the impact of adding the T1 score feature as an input vector on the performance of the models.

The bottom two plots are also scatter and box plots, with r on the y-axis and transformers used for embedding extraction on the x-axis. These plots are colored based on the application of the delta vector feature (yes/no). These plots help in understanding the impact of using the delta vector as input for the models compared to the concatenated T1 and T2 embeddings.

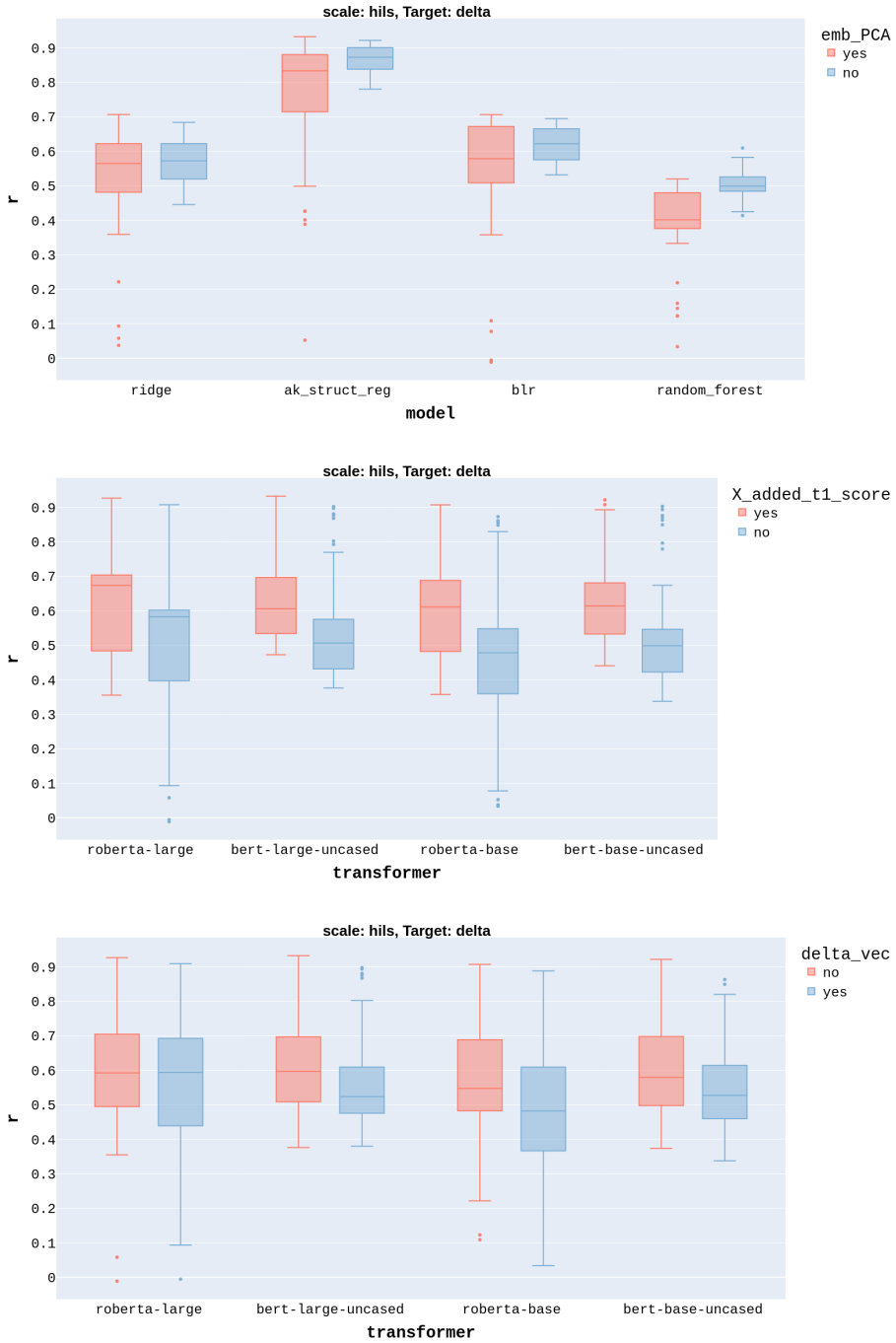

Figure 15: HILS delta prediction results. From top to bottom: Box plots grouped by regression model and PCA application, by transformer and inclusion of T1 score as a feature, and by transformer and use of delta vector embeddings.

model	mae	mse	rmse	r2	r	elapsed_time	n_features	transformer	delta_vec	PCA	pca_n	t1_feature
ak_struct_reg	0.794	2.497	1.085	0.884	0.933	1385.780	513	bert-large-uncased	no	yes	128	yes
ak_struct_reg	0.552	2.525	0.793	0.880	0.927	667.138	129	roberta-large	no	yes	32	yes
ak_struct_reg	0.874	2.827	1.165	0.869	0.922	2177.325	1025	bert-base-uncased	no	yes	256	yes
ak_struct_reg	1.173	3.288	1.561	0.856	0.922	10635.708	4097	bert-large-uncased	no	no	2	yes
ak_struct_reg	0.839	2.859	1.176	0.867	0.920	1835.615	1025	roberta-large	no	yes	256	yes
ak_struct_reg	1.193	3.457	1.538	0.845	0.916	984.390	257	bert-large-uncased	no	yes	64	yes
ak_struct_reg	1.448	4.140	1.900	0.827	0.911	947.806	129	bert-large-uncased	no	yes	32	yes
ak_struct_reg	1.048	3.356	1.410	0.847	0.909	728.370	129	roberta-large	yes	yes	64	yes
ak_struct_reg	0.693	3.527	1.022	0.832	0.909	639.903	33	roberta-large	yes	yes	16	yes
ak_struct_reg	1.364	4.204	1.831	0.819	0.908	771.528	129	bert-base-uncased	no	yes	32	yes

Table 6: Top ten results for hils delta prediction sorted on Pearson's r-value. where ak_struct_reg is the models created with the AutoKeras StructuredDataRegressor package

Interpretation of HILS Delta Prediction

Interpretation – HILS delta prediction In HILS delta prediction, models that included the original T1 score as a feature performed best. Interestingly, the delta vector input showed only a minor improvement over the concatenated embeddings, especially when combined with PCA. Once again, AutoKeras dominated the top 10, demonstrating its strength in modeling subtle changes over time in psychological constructs.

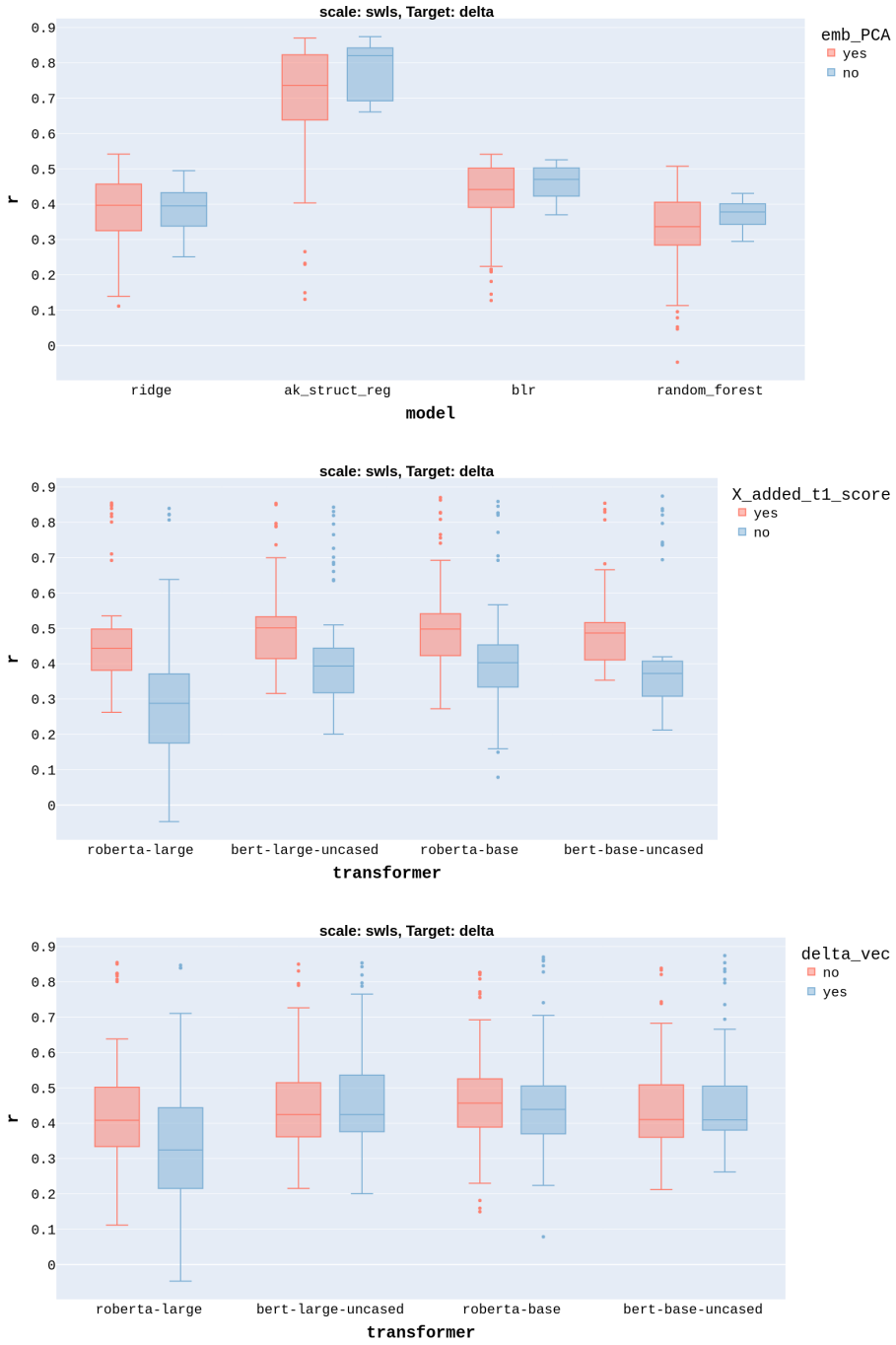

Figure 16: SWLS delta prediction results. From top to bottom: Box plots grouped by regression model and PCA application, by transformer and inclusion of T1 score as a feature, and by transformer and use of delta vector embeddings.

model	mae	mse	rmse	r2	r	elapsed_time	n_features	transformer	delta_vec	PCA	pca_n	t1_feature
ak_struct_reg	0.640	3.667	0.865	0.808	0.874	3242.494	1536	bert-base-uncased	yes	no	2	no
ak_struct_reg	0.637	3.695	0.863	0.807	0.870	747.152	65	roberta-base	yes	yes	32	yes
ak_struct_reg	1.370	4.531	1.862	0.768	0.863	1885.524	513	roberta-base	yes	yes	256	yes
ak_struct_reg	1.404	4.743	1.921	0.760	0.859	757.886	64	roberta-base	yes	yes	32	no
ak_struct_reg	1.450	5.022	1.967	0.746	0.855	11003.946	4097	roberta-large	no	no	2	yes
ak_struct_reg	0.976	4.240	1.351	0.780	0.854	1264.940	257	bert-base-uncased	yes	yes	128	yes
ak_struct_reg	0.888	3.936	1.240	0.795	0.854	3102.475	2049	bert-large-uncased	yes	no	2	yes
ak_struct_reg	1.370	4.860	1.825	0.752	0.851	1225.433	257	roberta-large	no	yes	64	yes
ak_struct_reg	1.072	4.144	1.451	0.785	0.850	2573.556	1025	bert-large-uncased	no	yes	256	yes
ak_struct_reg	1.510	5.114	1.888	0.738	0.847	649.854	65	roberta-large	yes	yes	32	yes

Table 7: Top ten results for swls delta prediction sorted on Pearson's r-value. where ak_struct_reg are the models created with the AutoKeras StructuredDataRegressor package

Interpretation of SWLS Delta Prediction

Interpretation – SWLS delta prediction In the SWLS delta prediction, using the delta vector was more beneficial than in the HILS case. The combination of delta vector, PCA, and T1 as input yielded the best results, again with $ak_{s}truct_{r}egmodelsdominatingtheleaderboard.Thissugg$ *relatedconstructs*.

0.4.3 Connection to Research Questions

Connection to research questions – HILS score The HILS results address RQ1 by demonstrating that psychological scores can be predicted from short text inputs using transformer embeddings. The strong r-values and consistent top performance of AutoKeras models support RQ2 and RQ4, indicating that neural architectures combined with dimensionality reduction can effectively model well-being constructs.

Connection to research questions – HILS Delta These results support RQ2 by showing that changes in harmony scores can be predicted using textual descriptions, especially when the T1 score is included. RQ3 is addressed through the observed effects of the delta vector and PCA preprocessing, and RQ4 is again supported by the superior performance of AutoKeras models.

Connection to research questions – SWLS score The SWLS score prediction results address RQ1 by showing that psychological well-being scores related to life satisfaction can be accurately predicted from short text inputs. The high r-values achieved by models using transformer-based embeddings, especially from bert-base-uncased and roberta-large, support this. RQ4 is supported by the superior performance of AutoKeras models compared to traditional regressors. The mixed effects of PCA suggest a nuanced impact of dimensionality reduction, contributing partially to RQ3.

Connection to research questions – SWLS Delta The SWLS delta results support RQ2 by demonstrating that semantic changes in word responses over time can predict changes in satisfaction with life. The combination of delta vector embeddings, PCA,

and the inclusion of the T1 score as a feature led to the highest predictive performance, directly addressing RQ3. Once again, the top results were achieved by AutoKeras models, reinforcing RQ4's conclusion that neural networks outperform traditional models in this task.

Overall summary of research questions Across both HILS and SWLS constructs, the results support RQ1 by confirming that psychological scale scores can be predicted from short, self-generated word inputs using transformer-based embeddings. RQ2 is addressed by the successful modeling of change scores (deltas), particularly when including the T1 scale score as a feature. RQ3 is supported through the measurable impact of preprocessing strategies such as PCA and delta vector calculation on model performance. Lastly, RQ4 is strongly validated by the consistent top performance of AutoKeras neural networks over traditional regression models, across both prediction tasks and constructs.

0.5 Discussion and outlook

In this section, we will discuss the implications and findings of the pipeline benchmark results for both score and delta prediction. Furthermore, we will provide ideas for future research and development of the pipeline

0.5.1 Discussion

PCA application

The methodology of this study involved applying Principal Component Analysis (PCA) to the entire dataset before cross-validation, which introduced a significant flaw - the risk of information leakage. This practice, while common for its simplicity, inadvertently allowed data from the validation set to influence the training process. The proper approach, fitting PCA within each fold of the cross-validation, maintains the independence of the validation data, ensuring a more accurate assessment of the model's generalization capabilities.

This methodological oversight was recognized post hoc, and a subsequent correction was made to the experiment design after the scope of this thesis. This correction aligns the methodology with standard practices, although the results within this document remain as derived from the initial approach.

Independent test set

The current study employed a 5-fold cross-validation approach for model evaluation, which, while effective for certain aspects of model assessment, does not encompass the use of an independent test set. This is a significant limitation in our methodology. An independent test set, a dataset that the model has never encountered during its training or validation phases, is crucial for a thorough and unbiased evaluation of the model's performance. Its absence in this study means that our results might not accurately reflect the model's capability to generalize to new, unseen data.

This oversight in our study design could lead to an overestimation of the model's accuracy and robustness. In future research, it would be beneficial to incorporate an independent test set, as this would offer a more comprehensive and realistic assessment of the model's performance. Including this phase would align the study more closely with the gold standard in machine learning model development, providing a more rigorous evaluation of the model's true predictive power.

The absence of an independent test set is a critical gap, especially when considering the potential for model overfitting or biases inherent in the training dataset. Without testing on new and independent data, it is challenging to confidently assert the model's efficacy in real-world scenarios. Therefore, while our 5-fold cross-validation provides some insight into the model's performance, the results must be interpreted with caution, acknowledging the

possibility of limited generalizability.

Comparable methodology

The differences in cross-validation methodology and dataset used in this study complicate direct comparisons with earlier studies such as [Kjell et al., 2022b] and [Kjell et al., 2021b]. While the previous studies employed a 10-fold cross-validation scheme, the current study utilized a 5-fold approach. Although both methods are standard, the 10-fold cross-validation is often considered more robust due to its larger number of folds, which typically provides a more precise estimate of the model's performance. The choice of a 5-fold cross-validation in this study was made to balance computational demand with performance estimation accuracy. Nevertheless, the variation in methodologies means that care must be taken when comparing the findings across these studies.

Furthermore, the use of a different dataset in this study also limits the comparability of the results with previous studies. The dataset used in earlier studies for score prediction may have different characteristics and properties from the dataset used in this study. This also limits the comparability of the findings, as the differences in the samples may impact the effectiveness and accuracy of the prediction models.

If this study had followed the same cross-validation methodology and used the same dataset for score prediction as previous studies, it would have been easier to compare and contrast the results. Similarly, for delta prediction, using the more robust 10-fold cross-validation approach would make the findings in this study more trustworthy. Since no earlier studies on word response construct delta prediction has been done it would set a better precedent following a standard from a similar task. This would increase the reliability and validity of the results and enhance our understanding of the effectiveness of prediction models for psychological constructs, both for score and delta prediction.

The challenges in comparing the results of this study with earlier studies, due to differences in cross-validation methodology and dataset used, could have been avoided with better planning, including a more structured literature study, and more experience with research projects that builds upon earlier work.

Transformer and PCA choice

For HILS score prediction, using different transformers won't make a drastic difference but *bert-base-uncased* and *roberta-large* has the best results. The application of PCA don't seem to have had a big impact on the accuracy of the model as can be seen in figure 13 and table 4. For SWLS score prediction PCA seems to be slightly less effective and the difference between transformer performance is slightly larger. Most of the results is otherwise much like for HILS score prediction as shown in figure 14 and table 5.

When it comes to delta prediction the result for both HILS and SWLS looks similar to

the results for score prediction. Above that it seems like using a *delta vector* was more effective when predicting SWLS deltas than HILS deltas. The *added T1* feature seems to have improved the result for delta prediction with both scales as seen in the figures 15 and 16 and tables 6 and 7.

Data set size

The size of the dataset is a crucial factor in the reliability and generalizability of study findings. With a small sample size, like in this study with only 477 participants, the results may not accurately reflect the true distribution of the population and may be subject to chance variations. It is therefore important to ensure that the sample size is appropriate for the research question and that it is representative of the population of interest.

Moreover, using larger and more diverse datasets can provide a better understanding of the relationship between psychological constructs and other variables of interest. This could be achieved through the collection of data from multiple sources, such as online surveys or clinical assessments. In addition, the use of standardized measures can facilitate comparison across studies and increase the reliability of the findings. The use of larger and more diverse datasets can provide more robust evidence to support the development of interventions and treatments for mental health conditions.

Data leakage

Data leakage is a critical issue in score prediction studies, and it is important to minimize its potential impact on the reliability of results. One source of bias due to data leaks is the use of concatenated T1 and T2 data points for score prediction to double the sample size to 954, effectively using two data points for each individual. When the T1 and T2 data points is used as separate samples, it may lead to overestimation of performance, thereby limiting the validity of the score prediction results. To address this limitation, future studies should use only one data point per individual and larger data sets to reduce the potential for bias and enhance the validity of the results.

Another limitation in this study is the application of PCA to the entire dataset before cross-validation. This approach may lead to data leakage in the pipeline and reduce the validity of the results. To ensure the validity of results in future studies, it is recommended to apply PCA separately on the training and testing datasets. In each fold of the cross-validation scheme, a PCA should be trained on the training set and then applied to both the training and test set. This would ensure that the principal components are based solely on the training data and not influenced by the testing data.

Alternatively, other dimensionality reduction techniques such as t-SNE[2] or UMAP[3] could be used instead of PCA. These methods are based on nonlinear mappings and may provide better representations of the data structure without the risk of data leakage. Thus, incorporating these techniques in future studies could lead to more accurate and reliable score prediction results.

In conclusion, data leakage is a crucial issue that needs to be addressed to enhance the reliability and validity of score prediction studies. To mitigate the potential for bias, future studies should consider using larger datasets or using only one of the two forms when T1 and T2 forms are completed by the same individual. Additionally, PCA should be applied separately on the training and testing datasets, or alternative dimensionality reduction techniques should be used to avoid data leakage in the pipeline.

Autokeras superiority

This study has shown that the AutoKeras StructuredDataRegressor is superior in performance, compared to the other regression models in this study, for score and delta prediction. In many cases, the r-values for the AutoKeras model exceeded 0.9, while no other model ever exceeded 0.8. Although the exact reasons for the performance advantage are unclear, it is possible that AutoKeras was better able to exploit the potential data leakage in our pipeline.

[2] t-SNE (t-Distributed Stochastic Neighbor Embedding) is a technique used for visualizing high-dimensional data in a lower-dimensional space, typically two or three dimensions. It works by modeling the similarity between pairs of data points in the high-dimensional space and then optimizing a low-dimensional embedding that preserves these similarities.

[3] UMAP (Uniform Manifold Approximation and Projection) is a dimensionality reduction technique that uses an exponential probability distribution in high dimensions, and any distance metric can be plugged in instead of Euclidean distances.

Additionally, the neural network outperformed all other models in each category, suggesting that it has the potential to achieve higher accuracy and reliability in these predictions.

While the results of this study are promising, further studies with larger and more diverse datasets, proper PCA application, and 10-fold cross-validation should be conducted to confirm these findings. The use of the same dataset used in a previous report [Kjell et al., 2022b] for score prediction would allow for proper comparison and validation of the AutoKeras model's performance.

If the superior performance of the AutoKeras model is confirmed through further studies, it could have significant implications for the field of psychology and mental health. Accurately predicting psychological constructs such as life satisfaction or harmony in life using AutoKeras could allow for the development of more effective interventions and treatments. This could lead to improved mental health outcomes and a better quality of life for individuals.

Moreover, the use of neural networks like AutoKeras could allow for more complex training data to be used in prediction models. This is because neural networks can handle multiple separate inputs with different formats, which is not as feasible with traditional regression models. This could allow for more comprehensive assessments of individuals' psychological states and better-tailored interventions. The findings from this study suggest that neural networks have the potential to revolutionize the field of psychology and mental health research.

Limited hardware

The benchmarks were conducted on very limited hardware (table 8), however, due to the large number of models and the size of the dataset, a benchmarking of this size should ideally be done on a server. The total time taken for all the benchmarks was approximately one week (164 hours and 36 minutes), excluding runs with errors. With error runs, the time taken was about three times as long in this study. Although it is possible to conduct the predictions on a personal computer, using a powerful server or scalable cloud GPU's would significantly reduce the time taken to complete the benchmarking process. This would enable researchers and practitioners to efficiently evaluate the performance of various models and potentially increase the accuracy of predictions for these constructs.

Challenges of using scales as a measure of constructs

Scales are widely used in social sciences and psychology to measure various constructs such as well-being, depression, anxiety, and personality traits. However, the use of scales as a single-dimensional scalar to represent a complex construct has been widely criticized. Most constructs are multidimensional, and reducing them to a single score may not accurately represent the construct's different dimensions [Fried, 2016]. For example, the construct of depression may have different dimensions such as negative affect, anhedonia, and somatic symptoms, and each of these dimensions may require a separate measurement. Combining

these dimensions into a single score may mask changes in the construct that occur at the level of individual dimensions. Therefore, the construct may have changed, some dimension going down and some up, without showing in the result, making it difficult to accurately measure change in the construct over time.

Another challenge with scales is that there are often several scales measuring the same construct, and they may represent different dimensions of the construct space. Choosing the most appropriate scale for a particular study may lead to different results. Therefore, using scales as the ground truth for a construct may not provide a comprehensive understanding of the construct being studied.

Moreover, the use of scales as a measure of constructs assumes that the construct is stable and can be measured accurately at a single point in time. However, constructs such as well-being or depression are dynamic, and their dimensions may change over time. A single scalar value may not represent the time variance of the separate construct dimensions well. The dimensionality and time variance in visualized in Figure 17.

Despite these challenges, better estimates of constructs are being produced, such as the LEAD study performed at Lund University. This study aims to establish a better standard for measuring constructs using a standardized format with expert panels. The LEAD study is creating a method to produce datasets that will be more accurate to predict on, and using a dataset produced with the LEAD method to train models would be more interesting. However, the availability of these datasets is currently limited. Therefore, more research is needed to address these challenges and produce more accurate and reliable measures of constructs.

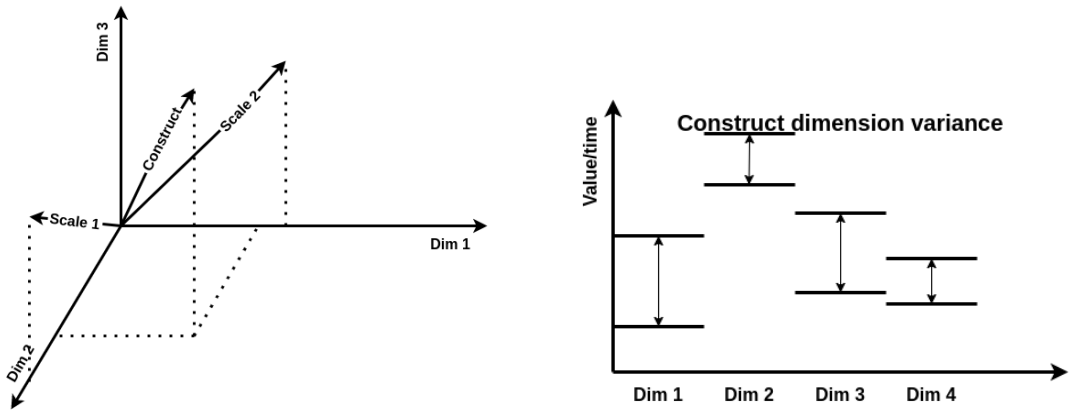

Figure 17: Scale predictions and construct value in the left picture and construct time variance per dimension in the right.

Overall, the use of scales to measure constructs presents several challenges that need to be addressed. Future research should focus on developing better methods for measuring constructs, including multidimensional measures and more comprehensive methods that take into account the construct's dynamic nature over time.

Limited number of constructs in this study

While the results of this study provide some insight into the performance of different machine learning models in predicting scale scores for two positive psychology constructs, "satisfaction with life" and "harmony in life," it is important to note that the study's scope was limited to these constructs. In reality, there are many more constructs and scales that researchers may want to use and predict scores for. It is possible that the performance of the models could differ depending on the complexity of the specific construct and the reliability of the scale being used. Future studies could expand on this work by testing a wider range of constructs from various fields and domains, such as mental health, personality, and education, among others.

By benchmarking more constructs, researchers could gain a better understanding of which models perform best across a range of different scenarios, and identify any patterns or factors that influence their performance. Furthermore, by exploring the limitations and challenges associated with predicting scale scores for a variety of constructs, researchers could identify areas where improvements are needed, and develop new techniques or methods to address these issues. Ultimately, this could help to increase the accuracy and reliability of predictions, and facilitate more effective research and clinical interventions in the future.

0.5.2 Research questions and answers

Improving Accuracy of Predicting Psychological Construct Scores Using NLP Compared to Earlier Methods The result of the benchmark demonstrates that selecting the appropriate transformer models, such as bert-base-uncased and roberta-large, can enhance the accuracy of predicting psychological construct scores from natural language data. Although the application of PCA had minimal impact on accuracy, the significant performance of AutoKeras StructuredDataRegressor indicates that switching techniques post-data collection can indeed improve prediction accuracy compared to earlier results and traditional regression methods.

Using NLP and Transformers with Regression Models to Predict Score Deltas from Text Data Our findings confirmed that NLP and transformers, combined with regression models, can accurately predict score deltas from text data. The use of a delta vector proved more effective for predicting SWLS deltas than HILS deltas. Additionally, incorporating the T1 feature enhanced delta prediction for both scales, validating the approach's efficacy.

Impact of Different NLP Preprocessing Techniques on Prediction Accuracy The choice of transformer model significantly affects prediction accuracy, with bert-base-uncased and roberta-large yielding the best results for HILS score prediction. PCA's impact was minimal for HILS scores and slightly less effective for SWLS scores. These findings suggest that while some preprocessing techniques like transformer selection are crucial, others like PCA may have a limited effect on prediction accuracy.However, PCA's ability to simplify calculations and reduce model size, while maintaining prediction accuracy, makes it a valuable preprocessing technique.

Effectiveness of Various Regression Models for Word Response Prediction Tasks Among the regression models evaluated, AutoKeras StructuredDataRegressor emerged as the most effective, consistently outperforming other models such as ridge regression, Bayesian ridge regression, random forest, and XGBoost. The AutoKeras model achieved r-values exceeding 0.9, whereas no other model surpassed 0.8, highlighting its superior performance for both score and delta prediction tasks.

0.5.3 Ethical aspects

Culturally biased dataset

The narrow scope of the dataset used in this study raises concerns about its generalizability to other cultural contexts. Since the data was collected through a website with users predominantly from the USA and India, cultural bias may be a potential criticism of the dataset. It is well-established that cultural differences can affect cognitive processes and the interpretation of assessment tasks, which could impact the validity of the study's findings. It is

therefore essential to consider cultural factors when interpreting and generalizing the results of this study.

To enhance the generalizability of the findings, future research should extend the analysis to different populations and cultural backgrounds. This will help establish if Autokeras is a useful tool for psychological construct prediction from word response data across different cultural contexts. Additionally, using a larger and more diverse dataset that covers a wider range of constructs can help to reduce the effects of bias and increase the generalizability of the findings.

0.5.4 Future ideas / outlook

AI image Likert scale

The use of AI generated images as a response format in a Likert scale survey is a novel concept that has the potential to offer new insights and advantages compared to traditional numerical response formats. With AI image generation, images could be created to better reflect the nuances of human emotions, making the survey more sensitive to subtle changes in attitude or opinion. Additionally, the use of images may reduce cultural biases and improve international comparability of results, as images can be more universally understood than numbers. However, it is important to note that AI generated images may not be universally accessible or understandable to all participants, and further research is needed to validate the reliability and validity of this response format.

AI for psych eval communication

The use of AI generated images and stories has the potential to revolutionize the way that patients communicate with healthcare professionals in a psychological setting. By providing patients with visual aids, AI has the ability to help patients express their thoughts and emotions in a more intuitive and accessible way. Furthermore, AI generated stories could be used as a way for patients to engage in imaginative scenarios that could help facilitate the therapeutic process. These advancements in AI technology have the potential to make psychological care more effective and accessible for a wider range of patients, especially those who struggle to express themselves through traditional methods. While further research is needed to fully understand the impact of these tools in a psychological setting, the potential for AI to positively impact mental health care is exciting.

Fine tuned transformer embeddings

In recent years, advancements in deep learning and artificial intelligence have made it possible to fine-tune transformer models for specific tasks, such as sentiment analysis or language generation. This opens up new possibilities for benchmarking, such as incorporating fine-tuned transformer models as a part of the prediction models, rather than extracting embeddings from pre-trained models. By incorporating fine-tuned models, it may be possible to obtain better performance and accuracy in predictions, as the models will have been trained to focus specifically on the task at hand. This is a promising area for future research and has the potential to lead to more sophisticated and accurate models in various applications, including the prediction of psychological constructs. However, the computational cost of fine-tuning these models can be high, and it is important to carefully consider the trade-off between improved performance and increased computational resources when designing experiments.

Tech nose data predictions

The use of electronic noses for construct prediction is a novel idea that has the potential to provide new insights and methods for predicting psychological constructs. Electronic noses are devices that mimic the olfactory system and have the ability to detect and distinguish

various odors. In a psychological setting, an electronic nose could be used to detect specific odors that are associated with specific emotions or mental states. This information could then be used to predict psychological constructs such as stress, anxiety, or mood. The use of electronic noses for construct prediction has not yet been tested, but might have the potential to provide a unique and innovative approach to understanding human behavior and emotions. Further research is needed to fully understand the potential of electronic noses in this context and to develop robust algorithms for predicting constructs based on olfactory data.

Using AI-Generated Images to Communicate Psychological Construct Scores and Deltas.

In addition to its potential benefits for patients and professionals, the use of AI-generated images to communicate psychological construct scores and deltas could also be particularly helpful for children and individuals with certain disabilities who may have difficulty understanding numerical or text-based information. Further future work could involve constructing a study to test the hypothesis that these images provide an accurate representation of the score or change in psychological construct scores compared to the perceptions of the participants in the study. Additionally, generative AI could be trained to better represent the perceptions of the participants and create more accurate images. As shown in Figure 18, a score can be represented by an image of a face with different emotions or expressions, while Figure 19 illustrates a delta between two scores with a side-by-side comparison of the two images, the images are generated from the same structured word responses used in this paper.

Figure 18: An image of four people describing their satisfaction with life with the words: unsatisfied displeased depressed upset troubled tired empty disappointed frustrated unhappy. With a score of 5 on the SWLS scale, the image is generated by the Midjourney [Midjourney,] image generator AI.

Note: The image generator (Midjourney) tended to produce male-presenting individuals by default, which may reflect embedded biases in generative AI systems., But could also be reflecting my bias choosing the images for this thesis

Representation of persons harmony in life at T1

Representation of persons harmony in life at T2

Figure 19: The upper image is of four people describing their harmony in life with: disorganized cluttered unmotivated unfocused disjointed uncertain unsure worried determined hardworking at T1. The lower image is of four people describing their harmony in life with the words: peaceful in-tune system happy integrated meaningful complementary loving loved sharing, instead. Going from a score on the HILS scale in the upper image of 11 to a score of 23 in the lower image, generated by Midjourney. [Midjourney,]

Video/audio question answering data predictions

The use of Video/Audio Question Answering (VAQA) data predictions is an emerging area in the field of construct prediction. With the advancement of technology, the use of whisper recognition systems such as speech-to-text technology has become more prevalent. The integration of speech-to-text technology, video analysis, and voice pitch analysis can provide a more comprehensive view of a person's mental state, behavior, and emotions. This can be useful in various fields such as psychology, sociology, and market research. In VAQA data prediction, the model takes in multimedia inputs and predicts a specific psychological construct based on the information obtained. This approach can provide more information and insights into the subject being analyzed compared to traditional methods, leading to more accurate predictions and a better understanding of the subject.

Using several layers of embeddings, CNN for prediction

It is possible to predict psychological constructs using a combination of deep learning techniques, specifically using Convolutional Neural Networks (CNNs) and Transformer models. In this approach, embeddings can be extracted from all the hidden layers of the Transformer and then formatted as images. These images can then be used as input for a CNN, which can be trained to predict the desired psychological construct. The advantage of this approach is that it utilizes the multi-layer representation capability of Transformer models and the ability of CNNs to identify patterns in images, providing a more comprehensive analysis. This approach has the potential to improve the accuracy and efficiency of construct predictions and may lead to new and innovative ways of exploring the relationship between input and target variables.

Concatenate embeddings for each word for prediction

It is possible to use all the embeddings for each word in a response for construct prediction, rather than using the mean embedding for the entire response. This approach can provide more detailed information about the language used in the response and how it relates to the psychological construct being measured. Neural networks, such as Artificial Neural Networks (ANNs) or Convolutional Neural Networks (CNNs), can be used to make predictions based on these word embeddings. The use of word embeddings instead of a mean embedding could result in more accurate predictions, as it takes into account the nuanced meaning of each word in the response. However, it may also increase the complexity of the prediction process and require more computational resources. Further research is needed to determine the efficacy of this approach and compare its performance to traditional methods.

One input layer for each embedding layer

Computational psychology researchers have been interested in developing machine learning models that can predict various psychological constructs based on data inputs for a long time. One of the challenges in this field is how to effectively represent and use the input data to make predictions. A possible solution to this challenge is to use parallel inputs instead of a long concatenated vector. In this approach, different features and inputs can be injected into

the model separately, allowing the model to process the information in parallel. For example, in the case of predicting delta changes, the words in the response could be inputted separately from the baseline score (t1 score). Additionally, other features such as voice pitch, video and text can be used as separate inputs to the model. This approach can potentially improve the accuracy and efficiency of the prediction process by allowing the model to better capture the complex relationships between the inputs.

Using AI-Generated Stories to Communicate Psychological Constructs

In the future, using AI-generated stories to communicate psychological construct scores and deltas to patients and professionals could be a valuable tool, particularly for children and individuals with certain disabilities. These stories could be used on an individual or group level and have the potential to effectively communicate complex psychological information in an engaging and understandable way. Future work should focus on conducting studies to test the hypothesis that these stories accurately represent scores and changes in psychological constructs compared to participants' perceptions. Additionally, generative AI models could be trained to better represent participants' perceptions. ChatGPT, as a language model, could be used for this purpose. One possible way to visualize the scores and deltas in a story format is shown in Figures 20 and 21, the stories are generated from the same structured word responses used in this paper..

There was once a person named John who found himself feeling constantly unsatisfied with his life. He was displeased with his job, depressed about his lack of social life, and upset about the state of his finances. He felt troubled by the uncertainty of his future, and tired from the constant struggle to make ends meet.

John had always been a dreamer, and he had imagined that by this point in his life, he would have achieved all of his goals. But as he looked around at his life, he couldn't help but feel empty and disappointed. He was frustrated with himself for not living up to his own expectations, and unhappy with the direction his life had taken.

John tried to make changes, but nothing seemed to stick. He would start a new job, only to feel unfulfilled and dissatisfied within a few weeks. He would try to meet new people, but found himself feeling awkward and out of place. No matter what he did, he couldn't seem to shake the feeling of unhappiness that had taken hold of him.

One day, John stumbled upon a quote that would change his life: "Happiness is not a destination, it's a journey." He realized that he had been so focused on achieving his goals that he had forgotten to enjoy the journey along the way. He had been so wrapped up in the idea of "success" that he had forgotten to appreciate the small things in life that brought him joy.

With this newfound perspective, John started to take small steps towards happiness. He started a gratitude journal, writing down three things he was grateful for every day. He made an effort to connect with people around him, even if it felt uncomfortable at first. He started to take walks in nature and to appreciate the beauty of the world around him.

Slowly but surely, John started to feel less empty and more content with his life. He realized that it wasn't about achieving some grand goal, but about finding joy in the everyday moments. He was still troubled by the uncertainty of the future, but he found comfort in the knowledge that he was making the most of his present.

In the end, John's journey towards satisfaction with life was a long and difficult one. But he learned that happiness isn't something that can be achieved overnight - it's something that is cultivated through the small actions we take every day. John had finally found a sense of peace and contentment that he had never known before, and he knew that he was exactly where he needed to be.

Figure 20: A story about a person describing its satisfaction with life with the words: unsatisfied displeased depressed upset troubled tired empty disappointed frustrated unhappy. With a score of 5 on the SWLS scale, generated by chatGTP.

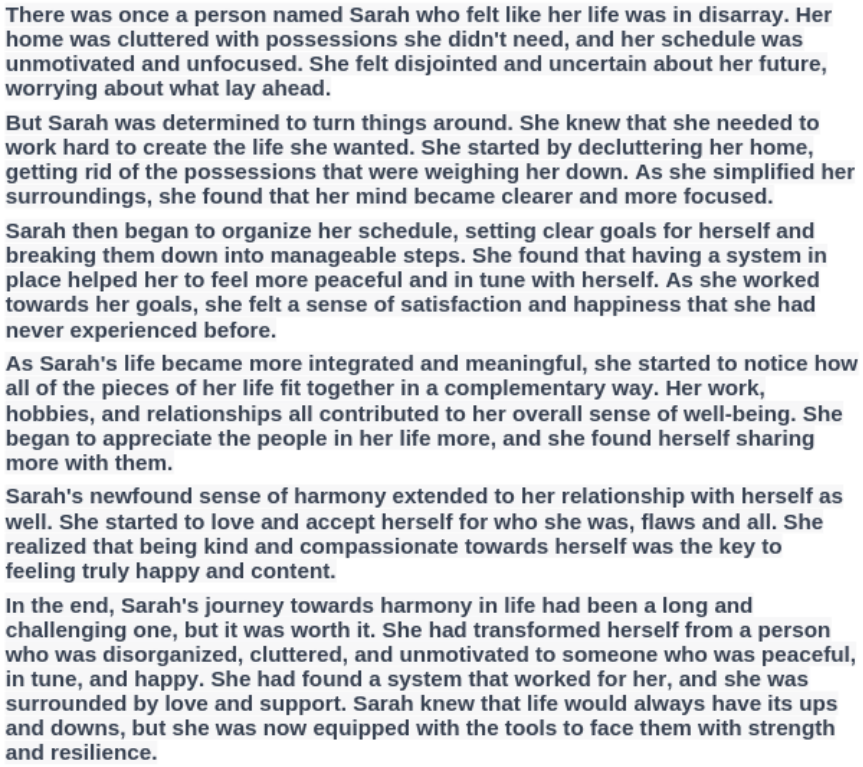

Figure 21: A story about a person going from describing its harmony in life with: disorganized cluttered unmotivated unfocused disjointed uncertain unsure worried determined hardworking, to describing its harmony in life with the words: peaceful in-tune system happy integrated meaningful complementary loving loved sharing, instead. Going from a score on the HILS scale of 11 to a score of 23, generated by chatGTP.

0.6 Appendix

0.6.1 Results and code

The rest of the results and the code is in the Github repository, https://github.com/Marcelunilsson/Master_Thesis.

0.6.2 Hardware

Component	Specification
Processor	Intel(R) Core(TM) i5-8600K CPU @ 3.60GHz
GPU	Nvidia Geforce GTX 970
RAM	16 Gb
SSD	Adata SP900

Table 8: Specifications for the hardware used in this study.

References

[Blampied, 2022] Blampied, N. M. (2022). Reliable change and the reliable change index: still useful after all these years? *The Cognitive Behaviour Therapist*, 15:e50.

[Dawes, 1982] Dawes, R. M. (1982). The robust beauty of improper linear models in decision making. In Kahneman, D., Slovic, P., and Tversky, A., editors, *Judgment under Uncertainty*, pages 391–407. Cambridge University Press, Cambridge.

[Devlin et al., 2018] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

[Diener et al., 1985] Diener, E., Emmons, R. A., Larsen, R. J., and Griffin, S. (1985). The satisfaction with life scale. *J. Pers. Assess.*, 49(1):71–75.

[Eichstaedt et al., 2018] Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D. A., and Schwartz, H. A. (2018). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

[Estrada et al., 2019] Estrada, E., Ferrer, E., and Pardo, A. (2019). Statistics for evaluating pre-post change: Relation between change in the distribution center and change in the individual scores. *Frontiers in Psychology*, 9.

[Fabrega, 1991] Fabrega, H. (1991). Psychiatric stigma in non-western societies. *Comprehensive Psychiatry*, 32(6):534–551.

[Fried, 2016] Fried, E. I. (2016). Are more responsive depression scales really superior depression scales? *J. Clin. Epidemiol.*, 77:4–6.

[Goldberg, 1970] Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychol. Bull.*, 73(6):422–432.

[Kjell et al., 2022a] Kjell, K., Johnsson, P., and Sikström, S. (2022a). Freely generated word responses analyzed with artificial intelligence predict self-reported symptoms of depression, anxiety, and worry. *Frontiers in Psychology*, 12.

[Kjell et al., 2015] Kjell, O., Daukantaitė, D., Hefferon, K., and Sikström, S. (2015). The harmony in life scale complements the satisfaction with life scale: Expanding the conceptualization of the cognitive component of subjective well-being. *Social Indicators Research*, 126.

[Kjell et al., 2021a] Kjell, O., Daukantaitė, D., and Sikström, S. (2021a). Computational language assessments of harmony in life — not satisfaction with life or rating scales — correlate with cooperative behaviors. *Frontiers in Psychology*, 12.

[Kjell and Diener, 2021] Kjell, O. N. E. and Diener, E. (2021). Abbreviated three-item versions of the satisfaction with life scale and the harmony in life scale yield as strong psychometric properties as the original scales. *J. Pers. Assess.*, 103(2):183–194.

[Kjell et al., 2019a] Kjell, O. N. E., Kjell, K., Garcia, D., and Sikström, S. (2019a). Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychol. Methods*, 24(1):92–115.

[Kjell et al., 2019b] Kjell, O. N. E., Kjell, K., Garcia, D., and Sikström, S. (2019b). Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychol. Methods*, 24(1):92–115.

[Kjell et al., 2023a] Kjell, O. N. E., Kjell, K., and Schwartz, H. A. (2023a). AI-based large language models are ready to transform psychological health assessment.

[Kjell et al., 2023b] Kjell, O. N. E., Kjell, K., and Schwartz, H. A. (2023b). AI-based large language models are ready to transform psychological health assessment.

[Kjell et al., 2022b] Kjell, O. N. E., Sikström, S., Kjell, K., and Schwartz, H. A. (2022b). Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Sci. Rep.*, 12(1):3918.

[Kjell et al., 2021b] Kjell, O. N. E., Sikström, S., Kjell, K., and Schwartz, H. A. (2021b). Natural language analyzed with ai-based transformers predict traditional well-being measures approaching the theoretical upper limits in accuracy.

[Lee et al., 2023] Lee, P., Fyffe, S., Son, M., Jia, Z., and Yao, Z. (2023). A paradigm shift from "human writing" to "machine generation" in personality test development: An application of state-of-the-art natural language processing. *J. Bus. Psychol.*, 38(1):163–190.

[Likert, 1932] Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

[Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

[Lovakov and Agadullina, 2021] Lovakov, A. and Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *Eur. J. Soc. Psychol.*, 51(3):485–504.

[McNemar and Meehl, 1955] McNemar, Q. and Meehl, P. E. (1955). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. *Am. J. Psychol.*, 68(3):510.

[Middel and van Sonderen, 2002] Middel, B. and van Sonderen, E. (2002). Statistical significant change versus relevant or important change in (quasi) experimental design: some conceptual and methodological problems in estimating magnitude of intervention-related change in health services research. *Int. J. Integr. Care*, 2(4):e15.

[Midjourney,] Midjourney (-). https://www.midjourney.com/.

[Robinson, 2014] Robinson, J. (2014). *Likert Scale*, pages 3620–3621. Springer Netherlands, Dordrecht.

[The Editors of Encyclopedia Britannica, 2022] The Editors of Encyclopedia Britannica (2022). *Student's t-test*.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

INSTITUTIONEN FÖR DATAVETENSKAP | LUNDS TEKNISKA HÖGSKOLA | PRESENTERAD 2025-06-11

EXAMENSARBETE Predicting Psychological Scale Scores and Deltas from Structured Word Responses
STUDENT Marcel Urrutia Nilsson
HANDLEDARE Dennis Medved(LU), Oskar Kjell (LU/Ablemind)
EXAMINATOR Maj Stenmark (LU)

AI som mäter ditt mående – med bara tio ord

POPULÄRVETENSKAPLIG SAMMANFATTNING **Marcel Urrutia Nilsson**

Kan du beskriva ditt välmående med bara tio ord? I detta examensarbete tränades och jämfördes flera AI-modeller för att tolka sådana ord – och resultaten visar att tekniken kan förutsäga både mående och förändringar i mående med relativt hög precision.

Psykologer har traditionellt använt långa formulär där människor sätter siffror på olika aspekter av sitt mående. Dessa summeras sedan till ett övergripande värde, en abstraktion som kan vara svår att relatera till, särskilt för personer med olika kognitiva förutsättningar. Formulären kan också upplevas som fyrkantiga och tidskrävande. I detta arbete utforskas ett alternativ, att låta människor beskriva sitt mående med egna ord, som sedan tolkas med hjälp av artificiell intelligens.

Studien bygger på data från 477 personer som vid två tillfällen fick beskriva sin livssituation med tio egna ord och samtidigt svarade på två etablerade formulär för att mäta välmående. För att tolka orden användes moderna språkmodeller, så kallade transformers, som översatte orden till sifferrepresentationer, ett slags koordinater (se Figur 1). Därefter tränades olika regressionsmodeller för att förutsäga deltagarnas formulärresultat. Den modell som presterade bäst, ett självoptimerande neuralt nätverk (AutoKeras), uppnådde en korrelation över 0,9, mycket högt inom experimentell psykologi.

Modellerna kunde även förutsäga *förändringar* i måendet. Genom att analysera hur ordens positioner förändrades mellan mättillfällena kunde AI:n avgöra om en person blivit mer harmonisk, lyckligare, eller tvärtom.

Sådan prediktion används redan i praktiken, bland annat i självhjälpsappar och internationella forskningssamarbeten. Med fler och mer varierade exempel att träna på skulle tekniken kunna bli ännu bättre på att förstå hur vi faktiskt mår.

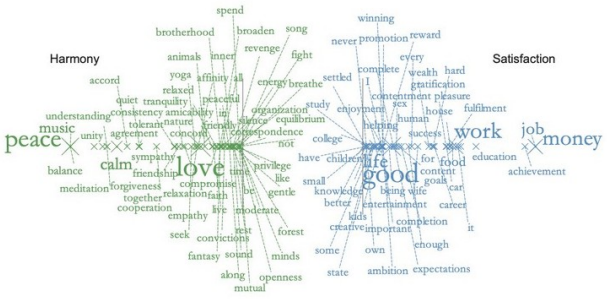

Figure 1: Exempel på ordrepresentationer

Resultaten är lovande, men fortsatt forskning krävs för att höja träffsäkerheten och säkerställa att metoden fungerar brett. Eftersom studien bygger på ett relativt litet datamaterial finns en risk att modellerna anpassar sig för mycket till just dessa exempel (s.k. överanpassning).

Att låta människor uttrycka sig med egna ord, i stället för att bara kryssa i rutor, kan bana väg för mer träffsäkra och tillgängliga sätt att mäta psykiskt välbefinnande. Med hjälp av AI öppnas nya möjligheter att förstå och följa mående, här genom att utgå från naturligt språk.