Automatic schema classification for Schema-Focused Therapy using *k*-Nearest Neighbour

Budi Han[1] , **Willem-Paul Brinkman**[1] , **Merijn Bruijnes**[1]

[1]TU Delft

Abstract

Personality disorders affect 1 in 7 adults reducing their quality of life. Schema-focused therapy (SFT) has become very popular in Psychotherapy in the treatment of personality disorders (PD), unfortunately there is still in increasing societal need for such mental healthcare. Automation in the assessment of SFT allows for Ecological Momentary Assessments (EMA). Resulting in a dynamic assessment of schema-modes and making the treatment more socially available. Automation is realised by Allaart in the form of a conversational agent (CA), but needs a better schema classification algorithm to improve its efficacy. The goal of this study is to evaluate the *k*-Nearest Neighbour (kNN) algorithm along with Allaart's dataset. The main question of the study is as follows: **How well can a schema be automatically classified from a textusing KNN?**. The method comprises of an experimental pipeline consisting of 4 stages: Labeling of dataset; pre-processing of the data; schema classification; and evaluation. kNN performed satisfactory in multi-label binary classification with a mean accuracy of 71% and a mean weighted f1-score of 0.62. kNN did not outperform other classification algorithm and performed inadequate in ordinal classification. Results indicate a contrast between majority and minority classes and found a recall of 100% on one of the majority classes. Hence, the data set is concluded to be imbalanced. Due to limitations on the dataset and the CA no reliable conclusion can be made on the performance of kNN in automated schema classification. This study proposed future research to conduct a field experiment such that the CA and its ability to perform EMA is evaluated and reliable data is produced.

1 Introduction

1.1 Motivation

Personality disorders (PD) have been found to affect 1 in 7 adults reducing the quality of life and resulting in high societal costs [1] [2]. Additionally, treatment for PD's are difficult to access; effective programs are scarce; and health professionals often lack training in the treatment of these disorders [3] [4].

Schema-Focused Therapy (SFT) is one program that has become increasingly popular in Psychotherapy for the treatment of personality disorders [5]. Despite the cost-effectiveness of the treatment compared to other treatments [6], there is still an increasing societal need for affordable mental healthcare and a general lack of providing such healthcare [7].

Schemas modes represent an unhealthy pattern of thoughts and behaviours that a patient's uses to cope with life, often brought about through childhood trauma [8]. Assessing and understanding a person's schema modes is done with the Short Schema Mode Index (SMI), a 118 item questionnaire which is scored using a 6-point scale ranging from "never or hardly ever" to "always". The items on the questionnaire are subsequently connected to the 14-factor model where it will relate to 1 of 14 schema modes.

However, SMI comes with its limitations. Firstly, SMI uses an extensive questionnaire which takes approximately 40 minutes to fill in [8]. Secondly, the therapist needs about 3-6 sessions for discussing the results to develop and establish a mode model for a patient. And lastly, schema modes are not constant, it represents a momentary state that can "flip" according to events and moods. Consequently, the SMI does not measure schema mode flips, resulting in only a single static measurement of what actually is a dynamic system of schema modes [9].

The vision of automating a part of the assessment process through a conversational agent (CA) was realized by Allaart [9]. The CA held a natural conversation in which it required a recent emotional story on initialization. The story was automatically analysed and followed up with evaluating questions. Consequently, the conversational agent allowed for multiple Ecological Momentary Assessments (EMA). Which gave the therapist more information and reduced the amount of sessions needed. Results show that the agent was a good predictor for schema modes and also hinted at a non-inferior user experience and time savings with using the agent [9].

Automation in the assessment will make SFT more dynamic through EMAs [9]. Treatment will be more accessible with the availability of the CA. Consequently, automation will result in a more time and cost-effective treatment and

make this therapy more socially available.

1.2 Problem

There are some limitations to Allaart's work. According to Allaart, the agent used is still lacking a text analysis algorithm and in its current state it could barely predict 2 of 7 schema modes. The paper argued that with a proper text analysis algorithm and a classification algorithm, this would improve the overall efficacy of predicting the schema modes [9].

Different classifiers exist [10]. However, due to the scope and time limit of this research only a subset will be taken into account. This research is centered around the performance of the *k*-Nearest Neighbour (kNN) algorithm. In later stages of the research kNN will be compared briefly with the two following classifiers: Support Vector Machine (SVM) and Recurrent Neural Networks (RNN).

No current research is conducted on the use of the kNN classifier in the assessment of PD in SFT. While research in multi-label classification for sentiment analysis exist, it cannot be compared to the classification of schemas [11]. Classifying patients into their representative schema carries a huge responsibility on the algorithm and the programmer. Since an algorithm tries to do a therapists job there is no room for error and therefor in depth analysis on the performance of such algorithms must be performed.

This research will thus answer the main question: **How well can a schema be automatically classified from a text using KNN?**

1.3 Approach

The rest of the research paper is organized as follows. Section 2 will mention related work and how it can benefit this research. Section 3 provides additional background information. Section 4 describes the method on how this research will answer the main question. Section 5 describes a detailed experimental pipeline and show the results of the experiment. Section 6 will provide a summary and an interpretation of the results. Section 7 and 8 will touch upon discussions, limitations, conclusions, contributions and future work. Responsible research will end the paper in section 9.

2 Related work

2.1 Sentiment analysis on Twitter Data

Research on kNN classification was conducted on sentiments analysis of Twitter data by Tyagi [11]. Sentiment analysis is the use of natural language processing to analyze and systematically identify subjective information. In essence, it is the processing of determining the emotional tone behind a series of words. This emotional tone can be used to gain understanding of attitudes, opinions and emotions expressed within textual data. In this study, performance of kNN was evaluated in polarity based classification. kNN classification is conducted with the use of the Bag-of-words method (BOW). This vectorization method turns documents into fixed-length vectors by simply counting the number of times a word appears in a specific document. The paper concluded that kNN was good classifier for classifying tweets into positive and negative with a accuracy of 79.3 % [11].

Tyagi's research investigated the classifiers under the framework of similarity-based learning using BOW. Instead, this research requires a deeper analysis of documents where it needs to be capable of capturing linguistic contexts of words rather than just similarity. Consequently, our strategy needs a different approach in vectorizing documents such that contexts is taken more into account. Also, our problem does not comprise of a polarity classification problem, where documents are either labelled as positive or negative. Instead, the schema focus therapy requires a multi-class, multi-label approach. It consists of 7 classes (schemas) where each document can be assigned to 1 or more classes. This study does give insights in what techniques to perform and introduces a structure for an experiment this paper can utilize.

2.2 Extracting schema modes from thought records

The identification of schema through thought records was conducted by Burger [12]. Burger's research mainly tried to answer if schema modes in thought record could be extracted by a machine. Her data set was collected from 320 participants resulting in 1600 thought records comprising of 5747 utterances. These utterances were than manually labelled with the help of a graduate student of clinical psychology. With this manually labelled data set Burger performed a benchmark with classification algorithms RNN, SVM and both regression and classification variants of kNN. Concluded from this benchmark was that per-schema RNNs performed best overall. Though it did not clearly outperform the other models. With these results the conclusion was made that schema modes could be extracted by a machine.

As compared to this reseach, the data set provided by Allaart will not be manually labelled. Labelling was done through the results from the SMI questionnaire that has been conducted with Allaart's CA. However, some conclusion can be drawn from Francisca's research that will contribute this research. kNN classification performed better than kNN regression, therefore kNN classification variant will be used for evaluating performance in this paper [12].

3 Background information

3.1 Allaart's dataset

The experimental dataset provided by Allaart's consists of stories and SMI questionnaire items. Stories represent a recent emotional event. This story reflects the participants feeling and emotions at specific moment [9]. The SMI questionnaire items focuses on the experiences and feelings of a subject over an extended time period [8]. The SMI questionnaire used in Allaart's data assesses 7 schema modes: Vulnerable, Angry, Impulsive, Happy, Detached, Punishing and Healthy. These schema modes are assessed through a number of psychological questions that are rated on a 6-point scale ranging from "never or hardly ever" to "always" [8]. According to Allaart, results of the SMI questionnaire can be used to classify the stories [9]. However, the raw dataset needs to be pre-processed first such that classification can be applied.

On receiving the dataset, There are two ways to classify schema modes. Binary classification is considered because

Allaart's dataset classifies stories into binary results by design. Binary classification results will mainly constitute in a baseline for evaluation and will serve in how adequate the study can be reproduced. Ordinal classification is considered because schema modes are realistically presented on a 6-point scale. This type of classification is more in line with the 6-point scale used in the SMI questionnaire and should indicate "how much" a patient is suffering from a PD rather than "yes" or "no".

3.2 Data pre-processing techniques

Natural language needs to be analyzed by the computer and real value needs to be extracted from text. Pre-processing a dataset involves pipe lining natural language into computer language. Different techniques exist and depend heavily on the task or problem to be solved.

Cleaning noise and formatting

The following techniques are required to clean the raw data set of its inconsistency and noisy raw data.

- Lower-casing
- Expand contractions
- Punctuation removal
- Stop-word removal
- Tokenization

Lower-casing ensures that same words with different representation are treated equally. Stop-word removal is another way of reducing the amount of distinct words in a document. It focuses on words that are in abundance and do not provide much information or value to the sentiment analysis. Both lower-casing and stop-word removal aid in reducing the vector space by dealing with outliers and similar words [13]. Hence, saving computing and time and efforts in processing large volume of texts [14].

Expanding contractions and punctuation removal ensures that the format of the data is clean and consistent, such that each word in a document can be examined without difficulty. Tokenization ensures that documents are represented as a list of words, such that it it will be easier to process.

Vectorization

For the machine to be able to process the data, natural language need to be transformed into a vector. Two techniques that are used to transform words into vectors are:

- TF-IDF (term frequency–inverse document frequency)
- Word-embedding

TF-IDf is based on a frequency count across all documents. It creates vectors be weighting the terms present in the document. If a term is present in all documents (word in abundance, thus meaningless), TF-IDF will assign it a low weight in the sentence vector. If a term does not appear in many other sentences it is important in identifying the sentence (carries more meaning). Hence, TF-IDF will give it a high weight. TF-IDf is a simple vectorization method and focuses on the similarity of documents. But, it does not account for semantic similarities in language.

Word-embedding tries to capture semantic relationship between words [15]. The main idea of word-embedding is that words that appear in the same context, have similar meanings. Word-embedding can learn feature presentations of words by two different methods. Skip-gram and CBOW (continuous-bag-of-words) are both 1-hidden-layer neural networks. In short, Skip-gram predicts the context by an input word. CBOW predicts the word by the context.

However words vectors can only capture to a limited extend when it comes to sentiment analysis. Paragraph vectors might by a more suitable approach examine multiple sentences [16]. Instead of learning feature presentations of words, the model can learn it for sentences and documents. In sentiment analysis word-embedding of documents is favored over TF-IDF since it accounts for semantics of words.

3.3 kNN

kNN has seen frequent use in classification problems because of its efficiency and simplicity [17]. Despite its efficiency, the kNN algorithm also has its disadvantages. In the upcoming sections the paper will briefly explain how the kNN works and will list some of its drawbacks.

Making predictions with kNN

The kNN algorithm is a similarity-based classification algorithm where similarity is based on distance measures. The closer the instances are to each other, the more similar. As the name of the algorithm already implies, it will look at the k number of neighbours closest to a new data sample. Each neighbour instance votes for their class and the class with the most votes is taken as the prediction class.

Distance measures

To determine the distance of the training dataset to the new sample the *Euclidean distance* is calculated. It takes the shortest distance between two points by calculating the square root of the sum of the squared difference between a new point and an existing point [18]. Another popular distance measures that needs to be considered is the *Manhattan distance*. Instead of the shortest distance it takes the sum of all the real distances between two points [18]. Both Euclidean and Manhattan are considered to be the simplest but most effective distance measures [19].

Drawbacks of kNN

kNN is a simple and intuitive algorithm, but comes with its drawbacks. The main points that this research needs to considers are the following:

1. **Curse of Dimensionality**

   As the number of input variables grows, the number of dimensions grows with it. This create a data space grows exponentially and makes the kNN hard to calculate distance measures [20] [21].

2. **Imbalanced data causes bias problem**

   kNN does not perform well on imbalanced data. Based on distance measures and density of the vector space model, the kNN will have a bias towards majority classes [20]. Consequently, minority classes will be poorly classified. [22]

3.4 Evaluation metrics

For the evaluation of the performance of the kNN algorithm in the classification of schemas, both binary and ordinal classifications need to be assessed. For different classifications, different metrics of evaluating results must be conducted.

Binary evaluation

In the evaluation of binary classification, accuracy is often used along other metrics such as precision and recall, since accuracy suffers from the accuracy paradox.

Accuracy is described as the ratio of the number of correct predictions to the total number of input samples. The accuracy paradox states that accuracy alone is not a good metric for predictive models [23] [24]. The underlying issue is that a class imbalance in the dataset can result in a simple model to have a high level of accuracy, which is too crude to be usefull. For example, in the incidence where class A is dominant, stated True in 99% of cases, then predicting everything is of class A has an accuracy 99%. Hence precision and recall are better measures [23].

In describing precision and recall it is necessary understand the difference between True positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) values. Both TP and TN describe cases in which the prediction class matches the actual class(True for TP and False for TN) FP describes cases in which the classifier predicted True but actual answer was False (Also known as Type I error). FN describes cases in which the classifier predicted False but was actual True (Also known as Type II error).

Precision (Positive Predictive value) is the ratio of correctly predicted True observations to the total predicted True observations. A high precision relates to low false positive rate. Precision is calculated as follows:

$$precision = \frac{TP}{TP + FP}$$

Recall (Sensitivity) is the ratio of correctly predicted True observations to all instances that are actually True. It is the fraction of instances that was actually labelled as True. Recall is calculated as follows:

$$recall = \frac{TP}{TP + TN}$$

F1 measure is the weighted average of recall and precision. It thus takes both FP and FN into account. Intuitively, the F1 measure is a better metric than accuracy. The f1-score is calculated as follows:

$$f1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

The confusion matrix is a tool that compares the actual values against the predicted values. It helps in visualizing the ratio between true/false positive/negatives values in a table, thus making direct comparisons easier to make.

The Receiver Operating Characteristic (ROC) curve is a graphical plot that illustrates the ability of a binary classifier with different thresholds [23]. It plots the True Positive Rate (TPR) on the y-axis against the False Positie Rate (FPR) on the x-axis. TPR is also know as recall and FPR is also know as the *probability of false alarm*. The area under the curve (AUC) indicates how good a model is good at making predictions. The higher the AUC, the higher the rate of TP and TN values [25].

Ordinal evaluation

Since results in the ordinal classification are ranked from 0 and 3, the Spearman correlation can be used. Where 0 means "does not fit the schema" and 3 means "Fits perfectly with the schema". The closer the predicted rank is to the actual rank the higher the correlation. The spearman correlation calculates a coeficient that ranges from -1 to 1. Where 1 means a perfect association of ranks, 0 means no association between ranks, and -1 means a perfect negative association of ranks.

4 Method

Based on related work and background the main question is further distilled in the following sub-questions:

How well can a schema be automatically classified from a text using KNN?

1. What data pre-processing techniques should we apply on the dataset to ensure optimal classification?
2. What is the most optimal KNN based algorithm for text classification
3. How well does KNN perform compared to SVM and RNN

To answer these questions this paper will conduct an experiment using Allaart's dataset and make an evaluation on the kNN algorithm.

4.1 Experimental method

The studies mentioned in the related work section provided this study with a structure for evaluating classifiers [11] [17] [12]. Firstly, data is collected or selected. Secondly, data preprocessing is performed to structure and vectorize their data. Thirdly, the respective classifier is constructed and trained on the training set. And lastly, classification is executed on testing set and evaluation is made.

This research will establish a similar structure that evaluates the kNN classification algorithm. Hence, the method of the experiment is structured as described in figure 1. A more detailed experimental pipeline will is provided in section 5.

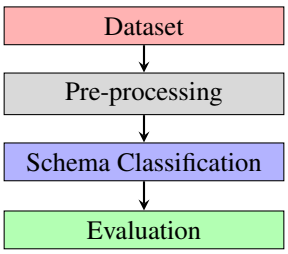

Figure 1: Experimental method

Dataset

This paper is based around the work provided by Allaart and his study on the conversational agent in the assessment of schema focused therapy. As a follow up research, his dataset will be used for experiments in the classification of schemas.

Pre-processing

After cleaning the dataset from it impurities, it must be structured and vectorized such that the classification algorithm can process it.

Schema Clasification

The kNN classifier is constructed and optimized for multi-label classification of schemas. The model is split of 80% training and 20% testing data.

Evaluation

The kNN algorithm will predict on the testing set. Results are evaluated and comparisons are made with SVM and RNN algorithms in order to find the best classifier for schema focused therapy. The performance of kNN will be examined and a conclusion is made on the research question.

5 Experiment and Evaluation

This section describes how the experimental pipeline is constructed and what tools are being used. The experimental pipeline is a detailed description of the method proposed in section 4. Please refer to Appendix A for the experimental pipeline.

Binary labelling

Binary labelling assigns either True or False to a specific schema. Allaart proposed a ruleset for labelling a specific schema mode. His ruleset is as follows: A schema mode is considered confirmed if any of the questions are answered with a 5 or 6, or if the average of the answers related to that schema mode is at least a 3.5 (See 1).

$$Schema = \begin{cases} True, & \text{if any item answered with 5 or 6} \\ True, & \text{if mean of schema items} \geq 3.5 \\ False, & \text{otherwise.} \end{cases} \tag{1}$$

Ordinal labelling

Ordinal labelling assigns a rank on a scale from 0-3 to a schema: 0 meaning "does not fit the schema" and 3 meaning "Fits perfectly with the schema". The following rule set has been applied. First the mean is calculated followed by a mapping on the scale of 0-3. Note that items answered with 5 or 6 always have a mean of at least 3.5 (See 2).

$$mean = \begin{cases} 3.5, & \text{if any item answered with 5 or 6} \\ & \text{AND mean} \leq 3.5 \\ mean, & \text{otherwise we take the mean} \end{cases} \tag{2}$$

Next, the mean is used to map documents on a scale from 0-3 (See 3).

$$Schema(mean) = \begin{cases} 0, & \text{if mean} \leq 3 \\ 1, & \text{if } 3 < \text{mean} \leq 4 \\ 2, & \text{if } 4 < \text{mean} \leq 5 \\ 3, & \text{if } 5 < \text{mean} \leq 6 \end{cases} \tag{3}$$

5.1 Pre-processing

Data trimming was applied to cleanse the dataset from impurities. To provide full transparency, the ruleset for this data manipulation is provided as follows:

1. Remove comments/questions unrelated to answering chatbot
2. Remove comments/questions that do not contribute to-classifying schema modes
3. Remove general responses (e.g. OK, Yes, No, Quit, Good Bye, Thank you)

Expanding contractions was performed with the *contractions* library. Lower casing and punctuation removal can all be performed with inbuilt functions from the *Natural Language Toolkit* (NLTK). From NLTK the *English stop-word corpus* was used to perform stop-word removal. Furthermore, the *NLTK's Tokenizer* module was used to tokenize Strings of words. *Gensim's doc2vec model* was utilized to perform word-embedding.

5.2 Schema classification using kNN

In classifying schemas *Scikit's learn kNN module* was used. This module allows the user to control the parameters for *distance measures*, *weight options* and the value *k*.

kNN Optimization

The kNN classifier was optimized for both binary and ordinal classification using *Scikit's learn Grid4Search module*. This module exhaustively generated the best candidates from a grid of parameter inputs.

In the case of kNN different parameters inputs were chosen for *distance measures*, *weight option* and *k*. For distance measures, the Manhattan and the Euclidean distance were considered; for weight options Uniform and Weighted options were considered; and for *k*, *k* = [4, ... ,60] was considered. The value cannot be to small since this can cause the classifier to over fit the new data samples.

The optimization was run with 10-fold cross-validation to reduce overfitting and ensure there was no randomness in tuning the optimal parameters. After exhaustively running the optimizer the following candidates were the best results for the kNN classifiers (See table 1)

Table 1: Optimal parameters for kNN in binary and ordinal classification

Method	Weight	d- measure	k
Binary	Weighted	Manhattan	4
Ordinal	Weighted	Manhattan	4

Binary and Ordinal schema classification

In both binary and ordinal schema classification a split of 80% training data and a 20% testing data was chosen. The optimal kNN classifier was fit on the training data.

5.3 Evaluation

Binary evaluation

Scikit's learn metrics module was used primarily for binary evaluation. This module provides functionality for *plotting confusion matrices* and *ROC curves*, producing a *classification report* and *calculating AUCs*.

Table 2: Overview confusion matrices

	TP	TN	FP	FN	Accuracy
vulnerable	26	163	31	69	0.65
angry	42	122	58	67	0.57
impulsive	5	218	9	57	0.77
happy	181	17	62	29	0.69
detached	28	155	38	68	0.63
punishing	10	206	21	52	0.75
healthy	266	0	22	1	0.92
Mean					0.71

Table 2 gives an overview the confusion matrices. Appendix B provides a detailed visualisation of the confusion matrices used for table 2. For each schema the accuracy and the amount of TP, TN, FP and FN values were identified. The ratio between schemas indicated that Healthy is a majority class with an accuracy of 92%. The confusion matrix showed that True was predicted 288 times out of 289 cases. Impulsive on the other hand had an accuracy of 0.72% and was predicted false 275 times out of 289 cases. The mean accuracy was 71%. It should be noted that accuracy is a crude evaluation metric as mentioned before.

Table 3: Classification report binary kNN

schema	precision	recall	f1-score	support
vulnerable	0.46	0.27	0.34	95
angry	0.42	0.39	0.40	109
impulsive	0.36	0.08	0.13	62
happy	0.74	0.86	0.80	210
detached	0.42	0.29	0.35	96
punishing	0.32	0.16	0.22	62
healthy	0.92	1.00	0.96	267
micro avg	0.70	0.62	0.66	901
macro avg	0.52	0.44	0.46	901
weighted avg	0.64	0.62	0.62	901
samples avg	0.74	0.69	0.68	901

Table 3 shows the classification report including precision, recall and f1 measure of all schemas. It should be noted that schema modes Happy and Healthy both have high scores. With healthy having 100% recall. Meaning the ratio of predicted True labels were all predicted correctly with the actual values. This again proves the schema mode Healthy to be the majority class. It can be concluded that all schema modes except for Happy and Healthy have relatively low precision, recall and f1-scores.

Figure 3 shows the micro-averaged and macro-averaged ROC curve. The ROC curves are based on the individual ROC curves per schema. The graph shows a notable bigger AUC for the micro-average than the macro-average. Meaning kNN performed better in classifying majority classes, but had difficulty classifying minority classes [25].

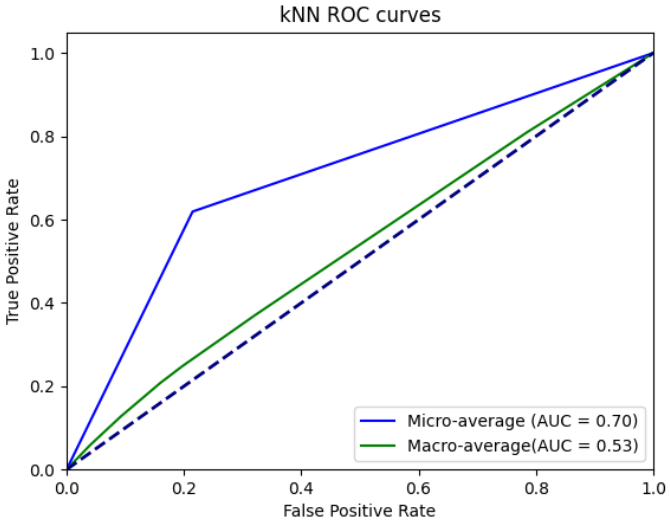

Figure 2: Micro/Macro averaged ROC Curve kNN

Ordinal evaluation

SciPi stats module offers a function that calculates a Spearman correlation coefficient. Table 4 shows the results of the coefficient per schema in ordinal classification. The table indicates that ordinal classification in general was inadequate. With the highest coefficient of 0.13 for the schema mode Vulnerable; lowest coefficient of 0.06 on the schema mode Healthy; and mean coefficient of 0.09 the conclusion can be made that the predicted values have no statistically significant correlation with the actual values. Meaning the kNN falls short in the classification of ranked/ordinal values.

Table 4: Spearman correlation coefficient per Schema

schema	Spearman
vulnerable	0.13
angry	0.08
impulsive	0.12
happy	0.06
detached	0.08
punishing	0.09
healthy	0.06
mean	0.09

Comparison with SVM and RNN
In comparing SVM, RNN and kNN the f1-score and ROC/AUC-curve will be utilized. In terms of f1-score, the RNN scored better overall with a schema average of 0.50 (See table 5). Though micro, macro, weighted and samples average indicate a similar performance between the classifiers.

Table 5: Classifier comparison f1-score

schema	SVM	kNN	RNN
vulnerable	0.27	0.34	0.38
angry	0.38	0.40	0.48
impulsive	0.07	0.13	0.17
happy	0.75	0.80	0.77
detached	0.18	0.35	0.36
punishing	0.20	0.22	0.34
healthy	0.93	0.96	0.95
schema avg	0.39	0.45	0.50
micro avg	0.63	0.66	0.66
macro avg	0.40	0.46	0.49
weighted avg	0.57	0.62	0.63
samples avg	0.66	0.68	0.66

Table 6 shows an overview of the AUC comparison between the classifiers averaged over all schemas. Appendix C is provided with detailed ROC-curves and corresponding AUC. In terms of the ROC-curve and the AUC, SVM scored best on the micro-average and RNN scored best on macro-average (See table 6). Meaning, SVM is better in classifying majority classses and RNN is better in classifying minority classes [25].

Table 6: Overview classifier comparison averaged AUC over all schemas

AUC	SVM	kNN	RNN
micro avg	**0.78**	0.70	0.69
macro avg	0.51	0.53	**0.54**

6 Results

Results summary
For binary classification kNN showed a mean accuracy of 71% with the majority class being the schema mode Healthy having a accuracy of 92%. The classification report opposed the accuracy paradox by providing precision, recall and f1-scores. The f1-score is the balance between precision and recall and should provide a better indicator than accuracy alone. Having a mean weighted f1-score of 0.62 with Healthy having the highest f1-score of 0.96.

For ordinal classification the kNN showed a mean spearman coefficient of 0.09. This low score indicated that the kNN performed poorly in predicting schemas on a scale from 0-3.

In comparing SVM, RNN and kNN, RNN scored best based on f1-score with a schema avg of 0.50. In terms of f1-score SVM scored best in the micro-average and RNN scored best in the macro-average. It should be noted that values are quite similar to each other and thus no real conclusion can be made on which classifiers performs best overall.

From the results, two main observations should be noted:

1. Schema mode Healthy has a recall of 100%
2. Macro-average is significantly lower than the Micro-average

Interpretation of the results
Based on two main observations made from the results a general interpretation can be made on the dataset provided with this study. With the schema mode Healthy being the majority class and having a recall of 100% a alarming indication should point towards an imbalanced data set. This indication is further supported by the Macro-average being significantly lower than the Micro-average. This implies that minority classes (E.g Impulsive and Punishing) are poorly classified, whereas your majority classes (E.g Healthy and Happy) are probably correctly classified. [25]. A good example of a minority class is the schema mode Impulsive.

Furthermore, the link between the stories and the SMI items have no real connection. Stories are based on a specific moment in time, whereas the SMI items reflect feelings and emotions over a period of time. It is hypothesised that the absence of a real connection between items and stories results in incorrect labeling. Consequently, the kNN has made its predictions based on unreliable data and evaluated poorly overall.

7 Discussions and limitations

7.1 Discussion

Comparison with related work
Compared to Burger's study, manually labelling might be preferred on Allaart's data set. Burger's study resulted in a significantly better Spearman Correlation on the performance of kNN. A mean score of 0.46 compared to a mean score of 0.09. However, manually labelling of a dataset would beat the purpose of automating the assessment in SFT.

In terms of pre-processing, emphasis is made on a vectorization method that considers sentimental value of documents. Thus, word-embedding was chosen over methods that are frequency based. For example, the BOW method that was used in the sentiment analysis on Twitter data [11]. Unfortunately, this decision did not reflect in the results of the kNN performance.

However, polarity based classification is rather rudimentary compared to schema classification. Since polarity requires a single class being either positive or negative and whereas the schema classification is a multi-label problem. It is safe to say that schema classification is rather a difficult tasks even with state-of-the-art word-embedding techniques.

Comparison with peer research
As part of the this research subject, additional studies have been conducted by peer students. Similar research has been conducted with particular focus on the performance of the SVM and RNN classifiers. These studies also concluded the data set provided by Allaart to be imbalanced [26] [27] [28].

Furthermore, a study on the generative aspect of the stories also indicated the dataset of having stories that are incorrectly labelled [29].

7.2 Limitations

The dataset suffered from imbalance and incorrect labelling. Consequently, the dataset did not provide a reliable foundation for training and testing a classification algorithm.

Firstly, the imbalanced dataset resulted in a biased classifier that was a poor predictor over minority classes and gave a false sense of accuracy over majority classes. Undersampling can be applied to reduce the amount of cases in the majority class to achieve balance. However, undersampling involves removing important information and reducing the size of the dataset. As the dataset gets smaller, the ability for a classifier to learn and predict will therefore decrease as well.

Secondly, the SMI questionnaire items, that were used to determine the story labels, did not correctly reflect the stories. The stories represent a momentarily mood or event of the participant. Whereas the SMI questionnaire items psychologically assessed participants over a certain time period. The stories that should represent a moment-to-moment assessment were thus labelled with items from a periodic assessment point of view. Hence, stories and the SMI items used for labelling had no real connection and were conceptually different [29]. (E.g Sad and Angry stories were being labelled as Happy and Healthy). For detailed examples of incorrect labelling see Appendix D.

8 Conclusions, contributions and future work

8.1 Conclusion

This study tried to answer the research question: **How well can a schema be automatically classified from a text using KNN?**. To answer this question, it was distilled into 3 sub-questions. These sub-question assisted in acquiring the experimental pipeline.

1) *What data pre-processing techniques should we apply on the dataset to ensure optimal classification?:* The main stages of pre-processing are cleaning noise and impurities; structuring and formatting documents; and computing a vector of every document. An important technique for psychological analysis is to maintain semantic relationships with word-embedding such as the Doc2Vec model.

2) *What is the most optimal KNN based algorithm for text classification?:* Based on Franzcisca's study the kNN classification is chosen for evaluation. In both Binary and Ordinal classifications a weighted kNN based on Manhattan distance with k=4 was preferred.

3) *How well does KNN perform compared to SVM and RNN?*: RNN scored best overall in terms of f1-score and was able to correctly classify minorty classes the best. SVM performed best in classifying majority classes. It must be noted that these metrics show similar results and thus no concise conclusion can be made on which classifier performs best in the classification of schemas overall.

To conclude, study has evaluated the kNN classification algorithm in the classification of schema modes. Allaart's data set was taken as starting point and the proposed experimental pipeline ensured a structured way for evaluation. kNN performed satisfactory in multi-label binary classification, but poorly in ordinal classification. Results show that kNN did not outperform other classification algorithms. The study also identified an imbalance in the dataset and cases of incorrect labeling. It must be acknowledged that these limitations on the dataset result in a unreliable foundation for evaluation. Hence, no reliable conclusion can be made on the overall performance of the kNN algorithm.

8.2 Contributions

The main contribution to the practical field is to point out that the CA used in Allaart's study not only needs a good text analysis algorithm and a good classification algorithm, but that the data collected needs to have real value. A true connection has to exist between stories and the SMI items such that labels are correctly identified. This in turn will provide a more reliable foundation for evaluating classifiers. This study also contributed in a experimental pipeline that can be utilized in the future to assess the performance of such classifiers. It should be noted that checking for imbalance should have been present in the early stages of the experiment.

8.3 Future work

To improve validity of the evaluation, future research should focus on providing more reliable data. Emphasis should be made on the real connection between story and SMI questionnaire. Since stories represent one moment in time, it would benefit future work by having more stories available per participant spanning multiple moments over a time period. Therefore, a field experiment is proposed utilizing the CA such that multiple stories can be captured. The CA should be incorporated in everyday life of participants and EMA should result in a more reliable dataset. Consequently, the aim of this future research will evaluate the CA and its ability to perform EMA. After collecting the data from the field experiment the real connection between stories and the SMI questionnaire items needs to be found. This will provide in trustworthy labeling, more data and overall a better foundation for evaluating classification algorithms.

9 Responsible Research

This section provides ethical aspects and scientific integrity of the research. Section 9.1 addresses the ethical aspects and considerations about automation in schema based therapy. Section 9.2 will justify data trimming of the data set acquired. Section 9.3 accommodates for full reproducibility of the research.

9.1 Ethical aspects

It should be noted that classifying algorithms in psychotherapy carry a huge responsibility towards patients with personality disorders. While automation can make treatments more socially available and provide auxiliary tools for assessment, the human therapist still remains a constant and essential part of psychotherapy. Therefore, automation in the field of SFT should be considered as a supplementary tool where final judgements should be made by the human therapist.

9.2 Research Data

As part of the pre-processing pipeline, data trimming has been applied. To justify the data manipulation, full transparency has been provided in the form a rule set in section 4.

9.3 Reproducibility

To accommodate for full reproducibility and verification of this study, this study provided the reader with a detailed experimental pipeline. The pipeline describes the order of the steps and which tools were utilized. Furthermore, the repository of the code will be made public on the following link: *https://github.com/budihan/Automatic-psychological-text-analyses-KNN*. It should be noted that Allaart's dataset will not be made public since this might raise privacy concerns.

References

[1] B. F. Grant, D. S. Hasin, F. S. Stinson, D. A. Dawson, S. P. Chou, W. J. Ruan, and R. P. Pickering, "Prevalence, correlates, and disability of personality disorders in the United States: results from the national epidemiologic survey on alcohol and related conditions," *The Journal of Clinical Psychiatry*, vol. 65, no. 7, pp. 948–958, 2004.

[2] S. Torgersen, E. Kringlen, and V. Cramer, "The Prevalence of Personality Disorders in a Community Sample," *Archives of General Psychiatry*, vol. 58, no. 6, p. 590, 2001.

[3] D. I. Soeteman, R. Verheul, and J. J. V. Busschbach, "The burden of disease in personality disorders: diagnosis-specific quality of life," *Journal of Personality Disorders*, vol. 22, no. 3, pp. 259–268, 2008.

[4] D. I. Soeteman, L. Hakkaart-van Roijen, R. Verheul, and J. J. V. Busschbach, "The economic burden of personality disorders in mental health care," *The Journal of Clinical Psychiatry*, vol. 69, no. 2, pp. 259–265, 2008.

[5] J. E. Young, J. S. Klosko, and M. E. Weishaar, *Schema therapy: A practitioner's guide*. Schema therapy: A practitioner's guide, New York, NY, US: Guilford Press, 2003. Pages: xii, 436.

[6] A. D. I. v. Asselt, C. D. Dirksen, A. Arntz, J. H. Giesen-Bloo, R. v. Dyck, P. Spinhoven, W. v. Tilburg, I. P. Kreimers, M. Nadort, and J. L. Severens, "Outpatient psychotherapy for borderline personality disorder: Cost-effectiveness of schema-focused therapy v. transference-focused psychotherapy," *The British Journal of Psychiatry*, vol. 192, no. 6, pp. 450–457, 2008. Publisher: Cambridge University Press.

[7] J. Lake and M. S. Turner, "Urgent Need for Improved Mental Health Care and a More Collaborative Model of Care," *The Permanente Journal*, vol. 21, 2017.

[8] J. Lobbestael, M. v. Vreeswijk, P. Spinhoven, E. Schouten, and A. Arntz, "Reliability and Validity of the Short Schema Mode Inventory (SMI)," *Behavioural and Cognitive Psychotherapy*, vol. 38, no. 4, pp. 437–458, 2010. Publisher: Cambridge University Press.

[9] D. Allaart, "Schema mode assessment through a conversational agent," *Delft University of Technology*, 2021.

[10] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, pp. 1093–1113, Dec. 2014.

[11] A. Tyagi and N. Sharma, "Sentiments Analysis of Twitter Data using K- Nearest Neighbour Classifier," *International Journal of Engineering Science and Computing*, vol. 8, no. 4, 2018.

[12] F. Burger, "Natural language processing for cognitive therapy: extracting schemas from thought records," *Delft University of Technology*, 2021.

[13] D. M. Asghar, A. Khan, S. Ahmad, and F. Kundi, "A Review of Feature Extraction in Sentiment Analysis," *Journal of Basic and Applied Research International*, vol. 4, pp. 181–186, Jan. 2014.

[14] C. Silva and B. Ribeiro, "The importance of stop word removal on recall values in text categorization," in *Proceedings of the International Joint Conference on Neural Networks, 2003.*, vol. 3, pp. 1661–1666 vol.3, July 2003. ISSN: 1098-7576.

[15] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *arXiv:1310.4546 [cs, stat]*, Oct. 2013. arXiv: 1310.4546.

[16] Q. V. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," *arXiv:1405.4053 [cs]*, May 2014. arXiv: 1405.4053.

[17] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "Using kNN model for automatic text categorization," *Soft Computing*, vol. 10, no. 5, pp. 423–430, 2006.

[18] H. S. Ranjitkar, "Comparison of A*, Euclidean and Manhattan distance using Influence map in MS. Pac-Man," *Faculty of Computing Blekinge Institute of Technology*, p. 61, 2016.

[19] Z. Li, Q. Ding, and W. Zhang, "A Comparative Study of Different Distances for Similarity Estimation," in *Intelligent Computing and Information Science* (R. Chen, ed.), Communications in Computer and Information Science, (Berlin, Heidelberg), pp. 483–488, Springer, 2011.

[20] M. Radovanović, A. Nanopoulos, and M. Ivanović, "On the existence of obstinate results in vector space models," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, (New York, NY, USA), pp. 186–193, Association for Computing Machinery, July 2010.

[21] F. Rahman, "k-Nearest Neighbors and the Curse of Dimensionality," Aug. 2020.

[22] Y. B. Wah, H. A. A. Rahman, H. He, and A. Bulgiba, "Handling imbalanced dataset using SVM and k-NN approach," (Johor Bahru, Malaysia), p. 020023, 2016.

[23] B. J. M. Abma, "Evaluation of requirements management tools with support for traceability-based change impact analysis," *University of Twente*, pp. 92–100, 2017.

[24] T. Afonja, "Accuracy Paradox," Dec. 2017.

[25] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, pp. 427–437, July 2009.

[26] I. Park, "Automatic Psychological Text Analysis using Support Vector Machine Classification," *Delft University of Technology*, 2021.

[27] J. O'Dwyer Wha Binda, "Active Learning In Reducing Human Labeling For Automatic Psychological Text Classification," *Delft University of Technology*, 2021.

[28] M. Zhang, "Automatic Psychological Text Analysis using Recurrent Neural Networks," *Delft University of Technology*, 2021.

[29] J. Lam, "Generative algorithms to improve mental health issue detection," *Delft University of Technology*, 2021.

A Experimental pipeline

Figure 3 shows the detailed pipeline of the experimental method proposed in section 4.

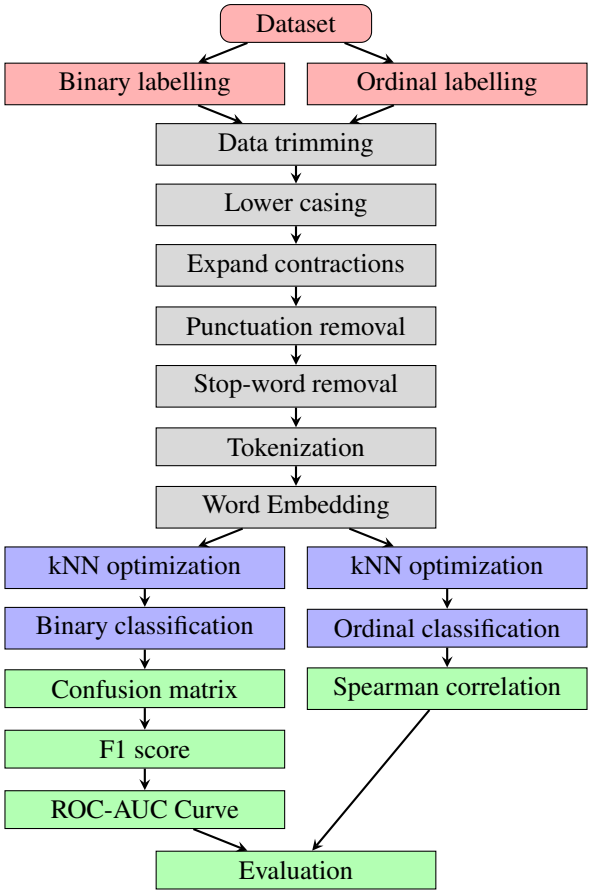

Figure 3: Experimental pipeline

B Confusion matrices

Figures 4 to 10 show a per schema confusion matrix. Values of these matrices are summarized in table 2: Overview of confusion matrices.

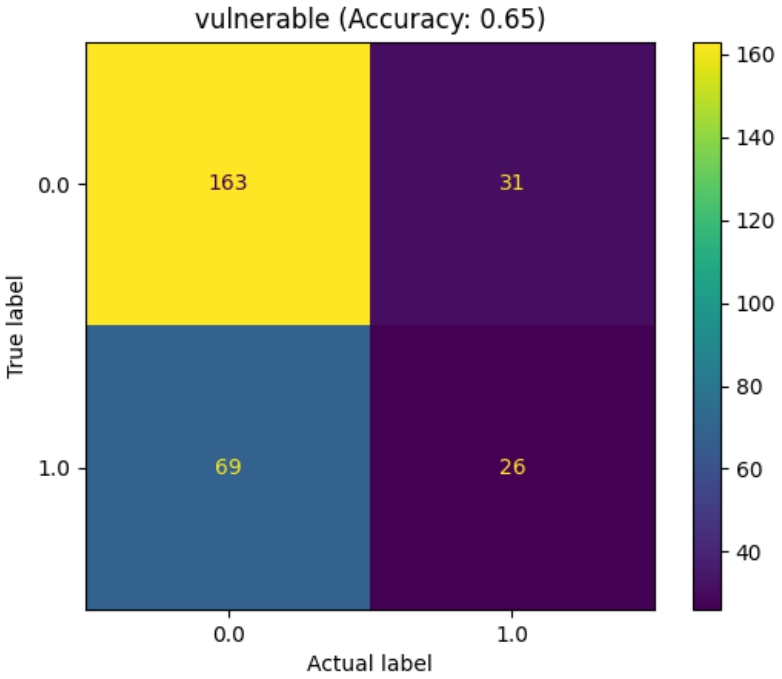

Figure 4: Confusion matrix Vulnerable

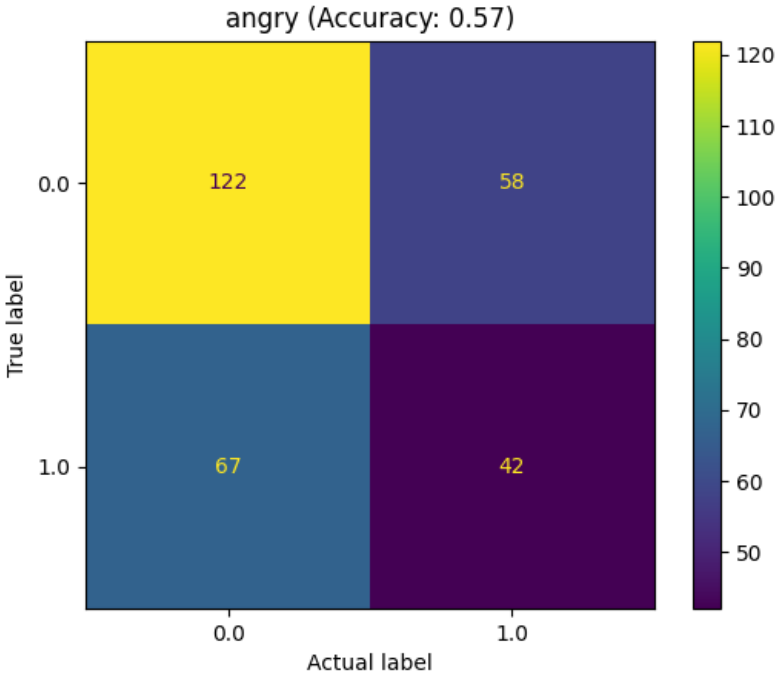

Figure 5: Confusion matrix Angry

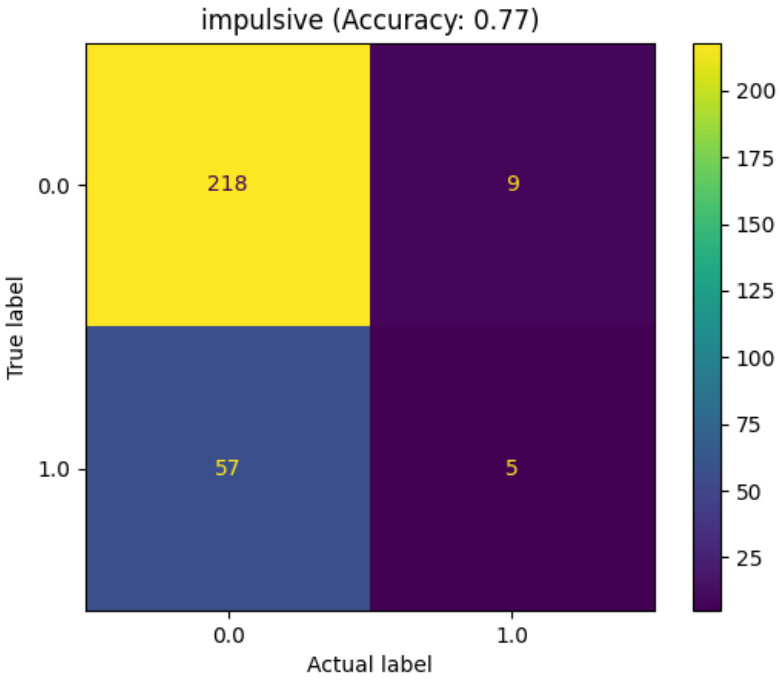

Figure 6: Confusion matrix Impulsive

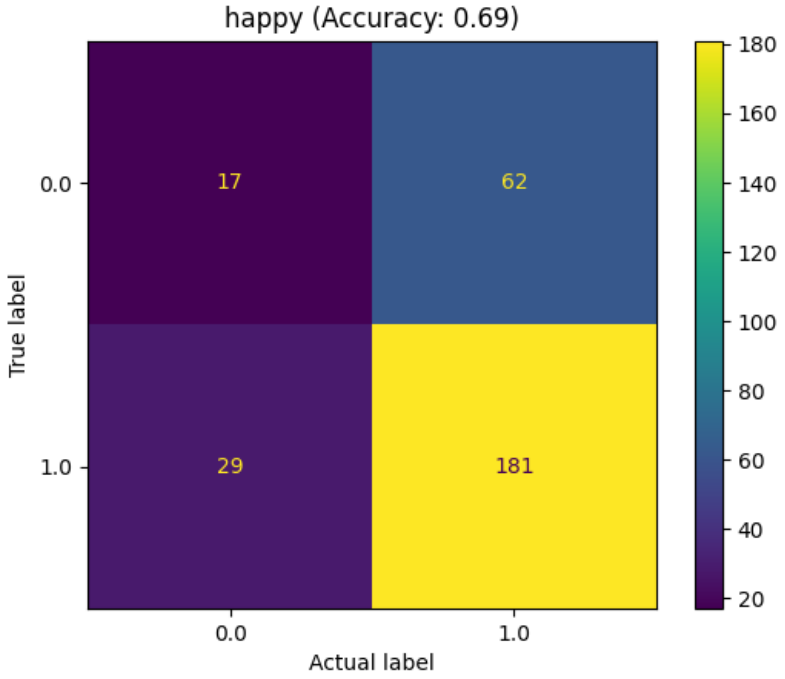

Figure 7: Confusion matrix Happy

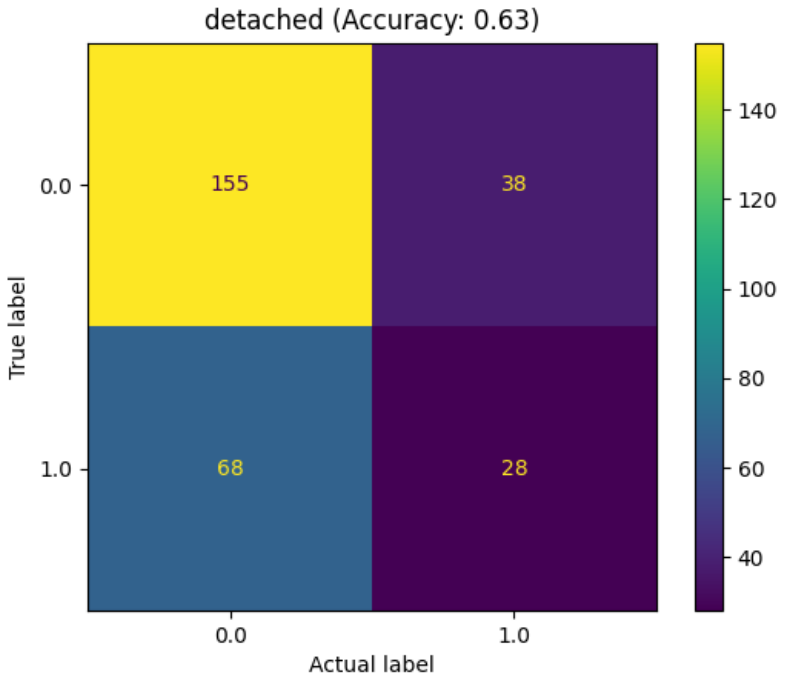

Figure 8: Confusion matrix Detached

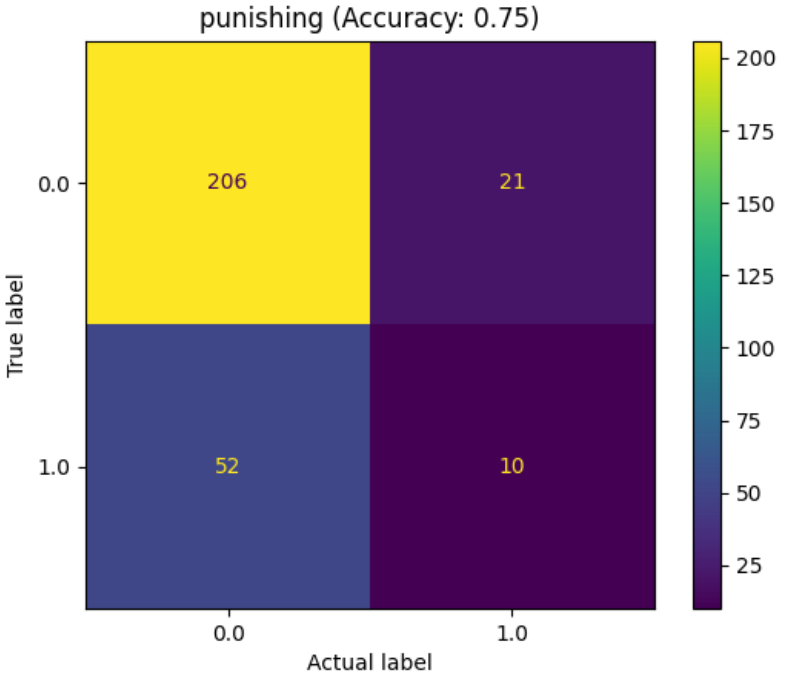

Figure 9: Confusion matrix Punishing

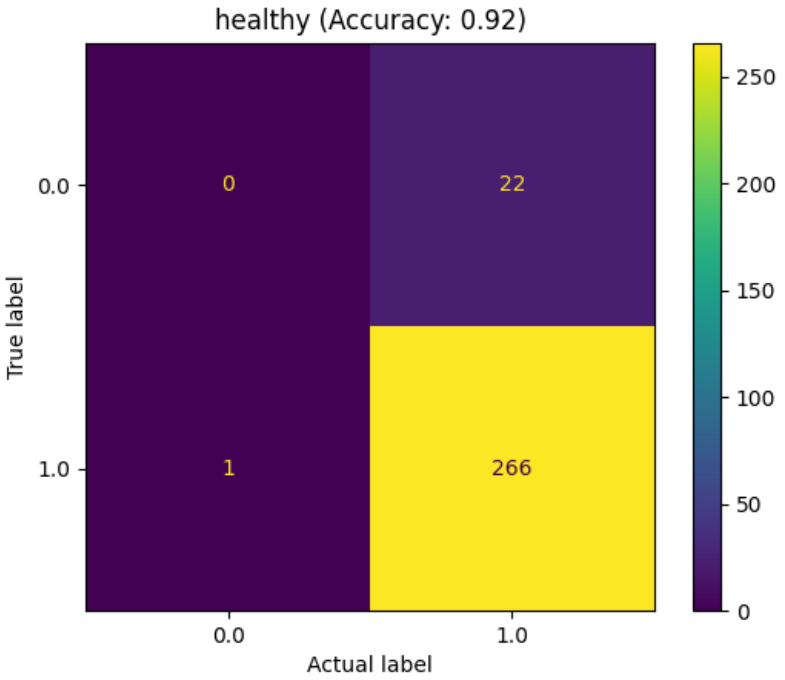

Figure 10: Confusion matrix Healthy

C ROC-curves and AUC for classifier comparison

Figures 11 and 12 show the micro and macro-average of the ROC curves and AUC of three classification algorithms. These results are summarized in table 6: Overview classifier comparison AUC.

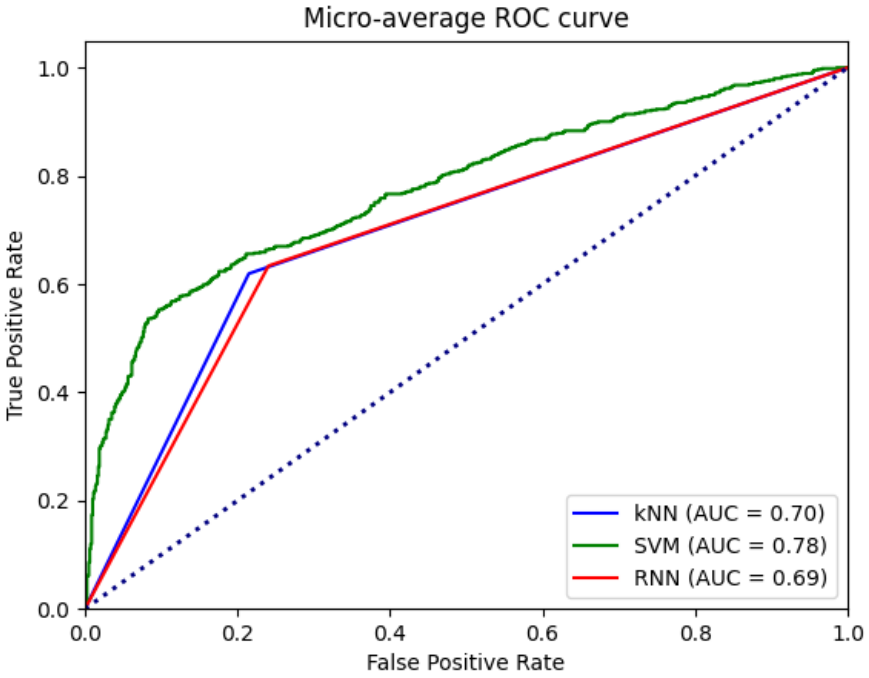

Figure 11: Plot ROC micro avg AUC

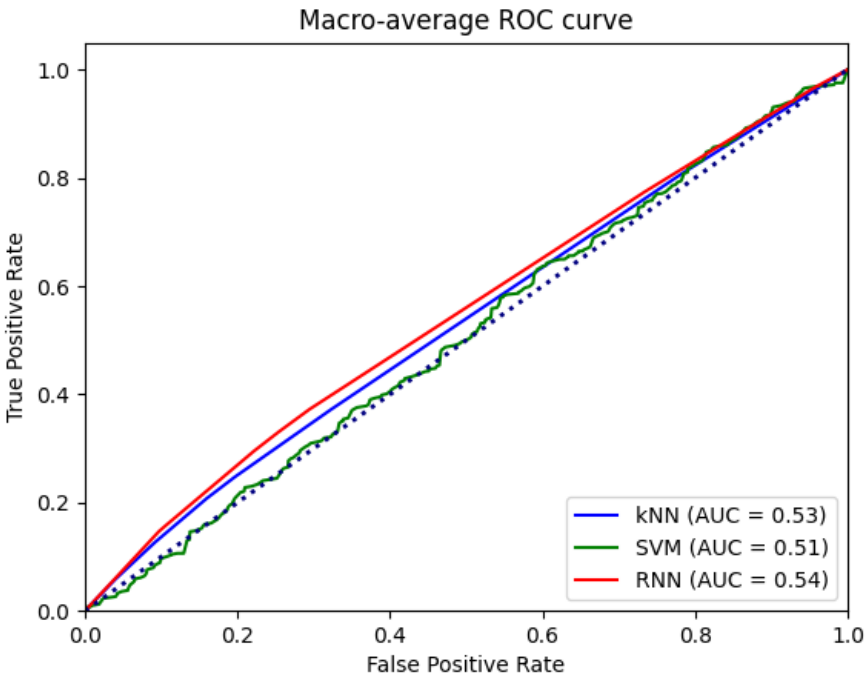

Figure 12: Plot ROC macro avg AUC

D Incorrect labeling

Figures 13 and 14 show examples of stories that were incorrectly labelled. Figure 13 shows a angry and stressful emotional event of a participant and figure 14 a sad and anxious story. However, both stories were labelled as happy by the SMI questionnaire.

Text						
hello, i recently had to shut my childcare business for the week. i was upset and angry as it meant i had to let down 6 families and refund their money. me and my boss lost money and it was a sad time for both of us. it was a stressful time and we had to contact a lot of people in order to find out if we could re open.						
is_vulnerable	**is_angry**	**is_impulsive**	**is_happy**	**is_detached**	**is_punishing**	**is_healthy**
FALSE	FALSE	FALSE	**TRUE**	FALSE	False	False

Figure 13: Incorrect labeling of story 1

Text						
i have not been feeling myself over the past few weeks and i have been finding myself not being able to complete tasks, i have been putting off a job that i said i would do for a friend that would help their business. i can't stop thinking about the fact that i haven't done it and it makes me feel really anxious and really bad but i can't bring myself to do it. i keep receiving emails from the friend but i have just felt so anxious i haven't been able to open them. i finally opened them and did the work and sent it back and i felt so relived but annoyed at myself for not doing it sooner and leaving it so long.						
is_vulnerable	**is_angry**	**is_impulsive**	**is_happy**	**is_detached**	**is_punishing**	**is_healthy**
FALSE	FALSE	FALSE	**TRUE**	FALSE	False	**TRUE**

Figure 14: Incorrect labeling of story 2