A Preliminary Study Examining an Automated Sentiment Analysis on Extracting Sentiment from Session Patient Records in an Eating Disorder Treatment Setting

Sophie Huisman

Faculty of Behavioural, Management and Social Sciences

Master Thesis Positive Clinical Psychology and Technology

First supervisor: S. de Vos

Second supervisor: dr. J. Kraiss

Abstract

Background: Clinicians collect therapy notes within session patient records which contain valuable information about patients' treatment progress. Sentiment analysis is a tool to extract emotional tones and states from text input and could be used for the evaluation of patients' sentiment during treatment over time. Therefore, this study aims to investigate the validity of an automated sentiment analysis on session patient records within an eating disorder (ED) treatment context against the performance of human raters.

Methods: A total of 460 patient session records from eight participants diagnosed with an ED were evaluated on its overall sentiment by an automated sentiment analysis and two human raters separately. The Inter-rater agreement (IRR) between the automated analysis and human raters and IRR among the human raters was analysed by calculating the intra-class correlation (ICC) under a continuous interpretation and weighted Cohen's Kappa under a categorical interpretation. Further, differences regarding positive and negative matches between the human raters and the automated analysis were examined in closer detail.

Results: The ICC showed a moderate automated-human agreement (ICC= .55) and the weighted Cohen's kappa showed a fair automated-human (k = .29) and substantial human-human agreement (k = .68) for the evaluation on overall sentiment. Further, the automated analysis lacks the inclusion of words specific to an ED context.

Discussion/Conclusion: This study suggests that the automated sentiment analysis performs worse in discerning sentiment from session patient records compared to human raters and cannot be used within practice. The automated analysis should be further investigated by including context-specific ED words and a more solid benchmark such as patients' mood should be established to compare the automated analysis to.

Keywords: Eating Disorders, Automated sentiment analysis, session patient records, Validation

Introduction

The term eating disorders (EDs) encompasses a range of serious psychological disorders characterized by disturbed eating patterns and can lead to (severe) somatic complications (American Psychiatric Association, 2013; Keel et al., 2012). The mean life-time prevalence of all EDs is 8.4% for women and 2.2% for men which has increased during the last decades (Bagaric et al., 2020; Galmiche and colleagues, 2019; Hoek, 2016). Further, EDs are considered to have the highest mortality rate amongst mental disorders caused by its complications (Arcelus et al., 2011), showing the serious consequences of this mental illness.

The three most common ED classifications are anorexia nervosa (AN), bulimia nervosa (BN) and binge-eating disorder (BED) (Keel et al., 2012). AN is characterized by extreme dietary restriction and underweight due to an intense fear of gaining weight and a distorted body image (American Psychiatric Association, 2013). In contrast, BN is characterized by recurrent episodes of uncontrolled and excessive food intake (binge-eating) which are often compensated for by purging or other unhealthy behaviours to prevent weight gain. In contrast, within BED the binge-eating episodes are not compensated for (American Psychiatric Association, 2013). Further, EDs that do not meet the criteria of one of the aforementioned disorders, but do create significant distress or functional impairment, are classified under the category of ‘other specified feeding and eating disorders’ (OSFED) (American Psychiatric Association, 2013).

Despite that there are several types of therapy available for EDs (Anderson et al, 2017), EDs remain difficult to treat and are accompanied with high levels of relapse, reflecting the often chronic nature of these disorders (Berends et al., 2018; Muzio et al., 2007; von Holle et al., 2007). Hence, it is important to further understand and monitor the recovery processes to protect individuals against relapse. One way to facilitate recovery is by monitoring the responsiveness of patients to treatment which can be supported by routine outcome monitoring (ROM) (Boswell et al., 2015). ROM is an instrument, using diagnostic indicators and severity scales, to periodically evaluate the patients' progress (de Beurs et al., 2011; Schulte-van Maaren, 2013). It can alert therapists when treatment does not show to be effective, indicate a worsening of symptoms, or reassure patients by providing insight into slight improvements within their situation (Youn et al., 2012).

However, ROM requires patients to fill out questionnaires about their own state, which may lead to subjective bias resulting in an over- or underestimation of the information (Karpen, 2018). Further, the ROM is supposed to be administered at fixed time intervals during treatment, which is burdensome for patients and time-consuming for therapists making it costly and not always feasible within a clinical setting (de Beurs et al. 2011; Gilbody et al., 2003; Norman et al., 2014). As a result, ROM is often only completed at the beginning and end of therapy, making for a less accurate representation of patients' treatment progress (Norman et al., 2014; Wampold, 2015). Therefore, the limitations of the ROM demonstrate that therapists could benefit from a less burdensome procedure and data usage to monitor patients' progress within treatment.

In fact, therapists already collect information about patients' treatment progress within session-by-session patient records (Swinkels et al., 2007). Session records are written texts by clinicians during therapy containing valuable information, such as patients' reactivity to and states during treatment, details of therapeutic conversation and clinicians' impressions of the patient (Maio, 2003; Percha, 2021). These records are an important part of treatment as they improve patient care by ensuring effective communication between clinicians and can support the substantiation of treatment choices (Ledbetter & Morgan, 2001; Patel, 2000). Evaluating these session records could also yield insightful information into patients' treatment process and progress in a less burdensome manner.

However, the usage of the session patient records in research is limited due to the records being long and complex, requiring more advanced and customised approaches to overcome the difficulties in extracting information from such texts (Berndt et al., 2015; Lee et al., 2020; Percha, 2021; Raja et al., 2008). The session records are classified as unstructured data, meaning that the qualitative texts are not stored into an organised predefined format, making it difficult to analyse with conventional analysis techniques (Boulton & Hammersly, 2006; Nikhil, 2015). One method to analyse texts is using human raters, however, this is a demanding and time-consuming task for researchers and is often not feasible when large amounts of text data are involved (Basit, 2003). Fortunately, throughout the last years new techniques have emerged supporting the analysis of unstructured text data in a more cost-effective and efficient manner (Smink et al., 2019). One such method is natural language processing (NLP) in which computer programs attain the ability to understand natural language in text or spoken words (Chowdhury, 2020). A subfield within NLP is automated sentiment analysis which aims to analyse natural language by using an algorithm operating through a set of rules to identify sentiment encompassing attitudes, emotions, appraisals, and the emotional tone within a text (Iliev et al., 2015). Hence, automated sentiment analysis could be particularly suited to analyse session patient record data since these contain emotional tones and states.

Different approaches exist for the execution of a sentiment analysis, a top-down and bottom-up approach. A bottom-up approach (unsupervised) uses unlabelled data through which the algorithm tries to uncover patterns within the data itself (Medhat et al, 2014; Priyavarat, 2017). A top-down approach (supervised) uses already labelled data, such as a predefined lexicon in which the polarity of sentiment-bearing bearing words are already pre-classified, in order for the automated analysis is able to predict the outcomes (Medhat et al, 2014; Priyavarat, 2017).

The analysis of sentiment has become increasingly popular and was mainly used for the mining of sentiment from online customer reviews, however, upcoming research started to examine the sentiment of patients' medical records (Hoerbst & Ammenwerth, 2010; Mäntylä et al., 2018). Despite this, the application of sentiment analysis within clinical practice remains limited, especially the sentiment within session patient records has not been widely examined.

Nevertheless, a few studies were executed within a clinical setting. An exploratory study by Provoost et al. (2019) executed an automated sentiment analysis on online behavioural therapy texts and found that the automated analysis performed similar to the human raters in discerning sentiment from the texts. Further, a study investigating the performance of four different sentiment analyses on healthcare-related texts against a human baseline found three sentiment analyses to have fair agreement and one to have moderate agreement with the human raters (Georgiou et al., 2015). Moreover, a study evaluating the sentiment on videos and comments related to AN found a fair agreement between the automated sentiment analysis and human raters (Oksanen et al., 2015). However, to date, only one study has investigated the performance of an automated sentiment analysis on written statements from patients diagnosed with AN regarding their body perception and exhibited the possibility of the extraction of sentiment from such statements (Spynczyk et al., 2018).

Despite these studies showing promising results, a challenge within this type of research is that there is no solid benchmark to compare the performance of the automated sentiment analysis to. For example, Provoost and colleagues (2019) suggested that the automated sentiment analysis performed as good as the human raters, however, the human raters showed a moderate agreement, meaning that they still differ in many cases regarding the rating of texts' sentiment. Hence, due to human raters' lack of consensus with one another, it cannot be determined with certainty whether the performance of the automated sentiment analysis is acceptable. Another point is that due to this research being conducted within the field of clinical psychology, thorough research is required on new technologies before they can be applied within practice (Ben-Zeev, 2017; Provoost et al., 2019). Hence, the automated sentiment analysis requires to be thoroughly researched and validated since there is little understanding on the application of an automated sentiment analysis within a clinical context. Further, automated sentiment analyses can be highly context specific as texts within different contexts may require different vocabulary and language, for instance, to analyse social media texts in contrast to clinical documents (Wilson et al., 2005). So, the automated analysis needs to be thoroughly examined on the use of its vocabulary within an ED context, as it may differ with the vocabulary used within other domains of mental health care.

In all, limited evidence exists on the performance of an automated sentiment analysis on session patient records within an ED treatment context and it is not clear whether the automated analysis can extract sentiment reliably and valid. The use of session records may be of added value as these texts are readily available to examine treatment progress without added burden for patients and clinicians and could be used on different texts related to an ED setting. Therefore, this study will examine how an existing automated sentiment analysis from 6Gorillas (6G) which uses a top-down lexicon-based approach with predefined lexicons, evaluates unstructured text data from session patient records in comparison to human raters. Further, due to the context-specificity of the automated sentiment analysis the differences in positive and negative matches regarding the rating of sentiment between the automated analysis and human raters will be examined in detail.

Methods

Design and procedure

The current study conducted a preliminary qualitative exploratory study in which a set of 460 session patient records were each evaluated on their sentiment by an automated sentiment analysis and separately by two human raters.

Data collection took place between February 2019 to April 2022 during which participants received outpatient treatment at a specialised ED treatment institution in the Netherlands named Human Concern (HC). Patients were diagnosed with an ED by a psychiatrist in collaborations with an intake team at HC. Participants visited their therapist once or twice a week for individual face-to-face treatment sessions which were also partly online via Jitsi (online video conference) due to the restrictions regarding the COVID-19 pandemic in the Netherlands, requiring patients and clinicians to meet online (Daemen, 2020). Therapy sessions concerned topics regarding recovery, autonomy and decrease of problematic eating behaviour by means of cognitive behavioural therapy and insight-giving therapy, patients received homework after the sessions to apply what they have learned in their daily lives (Human Concern, n.d). Further, at the start of treatment each patient received an account for an eHealth environment in which questionnaires and exercises were offered. Within this eHealth environment, patients were provided with a brochure explaining the aim of the research, were able to request an extended brochure and contact the researchers for further information and received an informed consent form which they could withdraw when they no longer wished to participate (see Appendix A and B).

The client advisory board of HC gave advice on the execution of the study regarding adherence to ethical principles regarding patient privacy and possible risk and harm and clarity of the brochure. The study was approved by the board of directors at HC and the Ethical Committee of the University of Twente (220422).

Participants

Participants were Dutch patients from HC with the criteria of having a minimum age of 17 at the time of providing informed consent and an ED diagnosis during data collection. A total of 149 patient from HC were asked to sign the consent form of which 12.1% rejected. A total of 131 patients provided consent of which a random selection was made including patients with different ED diagnoses with a minimum of forty session records.

In total the sample consisted out of eight patients including two patients diagnosed with AN, three patients with BN, one patient with BED and two patients with OSFED. Further, the study included five patients between the ages of 21-25, two patients between the ages of 26-30 and one patient between the ages of 31-35. The average duration of patients' treatment up to the start of the study was approximately 10 months (SD = 4.8).

Materials

Session Patient Record Data

The data utilised within the study were session patient record data provided by HC~~.~~ The session records were written by clinicians during treatment and include information from therapy sessions, treatment progression, ROM results and background information of the patient. However, not all the session records were suited for the analysis, as some only contained brief information about arranged appointments with other clinicians or institutions or descriptions of actions taken by the clinician(s) regarding administrative activities. Therefore, records that included one (or several) of the aforementioned actions or contained less than five words were excluded from the analysis by the human raters whereas the automated analysis only excluded records with less than five words or ones that did not include sentiment words.

Anonymisation

The model ‘deduce’, tailored to the Dutch language, was executed on the pseudonymised session patient records to anonymise the data (Menger et al., 2017). First, patient and postal codes, addresses, email addresses, telephone numbers, URLS and other contact information including those of relatives, clinicians from HC and other care providers and institutions were excluded. Second, the session records were tokenised, names and initials were changed to [NAME], dates to [DATE], dates indicating the start or end of treatment were transformed to a month and year, ages to [AGE] and locations or cities to [LOCATION]. However, the anonymity of the data could not be completely guaranteed and so, the board of directors of HC requested that the anonymised data could not be made publicly available. Consequently, the data was stored and archived within the first layer within the HC data archive, after the research was completed.

Further, the session records were accessed through a data platform (Snowflake) consisting out of different layers maintained by 6G and was conform to the Dutch legislations, ISO norms and the GDPR. The first layer was only accessible by the principal researcher at HC who first pseudonymised and anonymised the session records and after transferred them to the second layer in which all data analyses were performed. HC provided log-in credentials and a two-factor authentication to grant external researchers access to the second layer. After rating the session records an SPSS dataset was extracted from the first layer including only the sentiment scores of the automated sentiment analysis and human raters.

Automated sentiment analysis

To analyse the sentiment within the session patient records, an automated sentiment analysis from 6G tailored to the Dutch language and mental healthcare domain was used which was mostly manually validated but not in a scientific manner (6Gorillasn, n.d.). Before analysing the data, the sentiment analysis automatically pre-processed the data by transforming capital letters to lower case letters, removing stop words, numbers and words with only one character or underscores to improve the data mining functionality and prevent misleading results (Hemelatha et al., 2012).

Next, for the extraction of sentiment the automated sentiment analysis employed a top-down lexicon-based approach. Three lexicons were used , the primary lexicon was from NRC Word-Emotion Association containing English sentiment words translated into the Dutch language, a healthcare specific lexicon created by 6G and an adjustment dictionary from Ynformed (data science company) which changed or removed words with multiple meanings within a text (Royal HaskoningDHV, 2018). The lexicon awarded a positive or negative sentiment score to each individual sentiment-bearing words within a session record. Further, the automated sentiment analysis searched for words prior to a sentiment-bearing word to examine the semantic context using N-grams including bigrams (a two-word sequence) or trigrams (a three-word sequence). Consequently, the automated analysis could account for negations which reverse the polarity of a sentence (e.g., ‘not good’) and strengthening words (“extremely good”) (Dadvar et al., 2011; Farooq, 2017). The final sentiment score of a bigram was calculated by scoring the sentiment-bearing word with either ‘0’, ‘+1’ or ‘-1’ which was multiplied by two when the preceding word was a reinforcer and inverted when the preceding word was a negation. The final sentiment score was calculated by adding all the bigram scores of a session record divided by the total number of bigrams (6Gorillas, n.d.). The same approach was used for the trigrams, the last word of the sequence determined the sentiment and the two preceding words indicated whether the score was inverted or reinforced.

A final overall sentiment score was awarded to each session record which was an average of all the sentiment scores within a record-ranging between an interval of -1 and 1. Higher (positive) scores indicated greater positive sentiment, scores close to zero a neutral sentiment and lower (negative) scores indicated a greater negative sentiment of the record.

Human sentiment analysis

For the human sentiment analysis, the procedure of Provoost and colleagues (2019) was used as a guideline since this was the only study examinig the sentiment of texts within a Dutch clinical context.

The human sentiment analysis was divided into two parts, first, two human raters used a preliminary protocol to separately rate the first hundred session patient records. Every record was rated on a scale from one to seven, with '1' indicating very negative, '2' negative, '3' slightly negative, '4' neutral, '5' slightly positive, '6' positive and '7' very positive. After, a feedback session was arranged during which issues and difficulties with respect to the rating of sentiment were discussed upon which the protocol was revised. Hereafter, the new protocol was used for the evaluation of overall sentiment of the remaining session.

The category 'neutral' was assigned when a record was considered either objective (including no sentiment) or contained an equal number of positive and negative sentiment. Further, a separate category 'mixed' was used to indicate that a session record contained an equal number of positive and negative sentiment. Classifying both objective session records and records with an equal number of positive and negative sentiment as 'neutral' would be quite contradicting, as these two criteria were not equivalent in their meaning. However, since the automated sentiment analysis assigned the category 'neutral' to both criteria, the separate classification 'mixed' was created within the human sentiment analysis to explore the frequency of this phenomenon. Lastly, the category 'relevant' was used to indicate if a record did not fulfil the criteria for the analysis, further information about the criteria and the handled protocol can be found in Appendix C.

Data Analysis

Analyses were performed within the statistical program R (R Core Team, 2016) and Statistical Package of the Social Sciences (SPSS) 28 (IBM SPSS statistics) the alpha level was set at .05.

First, the human sentiment score was calculated by taking the average of both raters' sentiment score on each record and is referred to as the average human rating. Further, to be able to compare the outcomes between the automated and human sentiment analysis, the raw sentiment score on each session patient record were standardised for both analyses. Descriptives of the session records, category mixed and raw and standardised sentiment score of the automated and human analysis were provided, including the mean, standard deviation, minimum and maximum. A probability distribution for both the standardised automated and human sentiment analyses was created. Further, a scatterplot with a regression line was created to assess the strength of the linear relationship between the standardised sentiment scores of the average human ratings and the automated sentiment analysis. Lastly, a bivariate correlation with Perarson's coefficients was calculated defined as follows: 0 - .10 as negligible, .10 - .39 as weak, .40 - .69 as moderate .70 - .89 as strong and .90 to 1 as very strong (Schober, Boer, & Schwarte, 2018).

Further, for overall sentiment three categories were created for the raw and standardised sentiment scores of the automated sentiment analysis and human raters. For the automated analysis the raw sentiment scores were categorised as follows: negative for values smaller than -.01, positive for values larger than .01 and neutral for values between -.01 and .01 for a larger margin of values. For the standardised automated sentiment scores the categories were created as follows: negative for values smaller than -.03, positive for values larger than .11, due to the standardised scores not being equal to zero the category neutral was defined as values between the positive and negative category. For both human raters only the raw scores were used since the use of the standardised sentiment scores resulted in the same category distribution. For the human raters the category negative indicated values smaller than three, positive for values larger than three and neutral to values equal to three. Lastly, a contingency table was created including both human raters' raw sentiment scores with the frequency distributions of negative, neutral and positive scores between the human raters were displayed in order to obtain more insight into the relationship between the two variables.

Human-automated agreement

Categorical Interpretation. A weighted Cohen's kappa was calculated to assess the inter-rater agreement (IRR) which measured the extent that two (or more) examiners agree on their assessment decisions (Lange, 2011). The weighted Cohens kappa accounted for ordinal categorical data and was used to measure the polarity of a text in terms its direction (category). The weighted Cohen's kappa was calculated to examine the IRR between the standardised categories of the automated sentiment analysis and human raters (Cohen, 1960; Devitt & Ahmad, 2007). It was chosen to only use the first human rater's standardised sentiment scores since the average human sentiment scores were unlikely to be equal to the category neutral and reasonable categorical agreement (k = .68) between the human raters. Standardised catgeories for the first human raters were defined as follows: the median was classified as neutral, sentiment scores below the median were classified as 'negative' and above the median as 'positive'.

Values for the weighted Cohen's Kappa range between -1 and 1. For the interpretation, values below 0 indicated no, values between 0 and .20 none to slight, values between .21 to 0.4 fair, values between 0.41 to 0.6 moderate, values between 0.61 to 0.8 substantial and values between 0.81 and 1 almost perfect reliability (Landis and Koch, 1997).

Continuous interpretation. The Intra-class correlation (ICC) can be used to assess the IRR on continuous data and data with missing values (Devitt & Ahmad, 2007). The ICC correlated the sentiment scores on each session record with each other to measure the intensity of the agreement between the automated analysis' and human raters' on overall sentiment.

An ICC(2,k) was used, meaning that each text is rated by each rater, are the only raters of interest and that the average of the human raters' sentiment scores were used. Further, the ICC accounted for a two-way mixed effect model based on an absolute agreement, meaning that the human raters were the only raters of interest and assessed whether the two analyses assigned the same sentiment score to a text (Koo & Li, 2016). Further, for each patient the ICC was calculated between the first human rater and the automated analysis. An ICC(3,1) was used, meaning that each text is rated by one human rater of interest, accounting for for a two-way mixed effect model based on an absolute agreement (Koo & Li, 2016).

Values for the ICC ranged between zero and one and were interpreted as follows: values less than 0.5 indicating a poor, between 0.5 and 0.75 a moderate, between 0.75 and 0.9 a good and greater than 0.9 an excellent reliability (Koo & Li, 2016).

Human-human agreement

Categorical interpretation. A weighted Cohen's kappa was calculated to assess the IRR between the raw categorical scores of the human raters.

Continuous interpretations. The ICC was not calculated for the agreement between the human raters as there were only two raters using an ordinal scale which may lead to an expected overestimation of the agreement between the human raters (Denham, 2016). Standardisation could not account for this issue.

Qualitative analysis

In order to assess the differences between the automated and human sentiment analysis in more detail, a line graph was created for each patient which illustrated the sentiment score of a patient over time. The graphs included both the automated and average human ratings' standardised sentiment scores on each session record (y-axis) and the number of records (x-axis). Further, (large) differences between the automated and human sentiment analysis regarding sentiment scores were examined and reflected upon by comparing the sentiment-bearing positive and negative matches of the automated and human analysis. Accordingly, a wordlist was created for words specific to the ED context and different diagnoses which were not considered during the automated analysis but indeed during the human analysis. Lastly, positive and negative sentiment matches identified by the automated analysis which were not considered or considered of the opposite sentiment by the human raters, were listed into a table.

Results

Descriptive statistics

Patient session records

The total data set consisted out of 460 session patient records with an average of 57.50 (SD = 48.02) records per patient. The first human rater identified 268 (58.3%), the second rater 263 (57.1%) and the automated sesntiment analysis 315 (68.5%) records as relevant for the analysis. Within the human sentiment analysis 45 (9.8%) session records were categorised as ‘mixed’ which is 70.3% of the session records that were classified as ‘neutral’.

Continuous comparison between the human and automated sentiment analysis

A summary of the descriptive statistics regarding the overall sentiment on the session patient records of the average human rating and the automated analysis can be found within Table 1. The automated sentiment analysis shows a larger range of values for the standardised scores in comparison to the human analysis, as the human raters only used seven fixed scores. Further, Figure 1 shows a normal distribution for the automated analysis in which most of the data is centred around zero, however, it shows some outliers as well. Figure 2 shows a more varying normal distribution for the average human rating with less outliers. Lastly, Figure 3 shows a positive correlation between the sentiment scores of the automated analysis and average human ratings with a moderate Pearson’s correlation (r = .41, p < .001)

Table 1

Raw and standardised mean (M), standard deviation (SD), Minimum (Min) and Maximum (Max) of the overall sentiment from the human and automated sentiment analyses

	Average Human Rating (n = 263)				**Automated Analysis (n= 315)**			
	M	**SD**	**Min**	**Max**	**M**	**SD**	**Min**	**Max**
Raw scores	3.76	1.14	1.50	7.00	.05	.07	-.20	.50
Standardised scores	0.00	1.00	- 1.99	2.85	0.00	1.00	- 3.62	6.42

Figure 1

Histogram with normal curve for the standardised automated sentiment scores

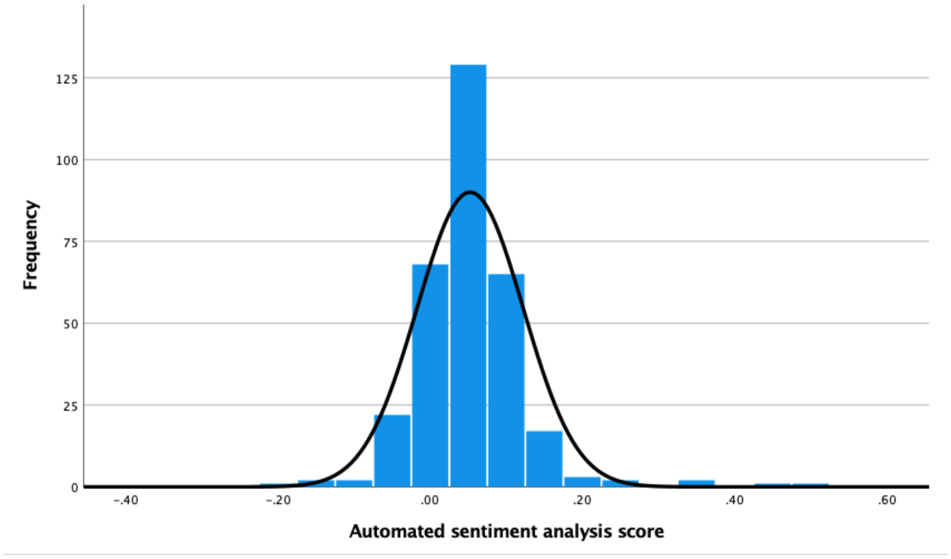


Figure 2

Histogram with normal curve for the standardised average human sentiment ratings

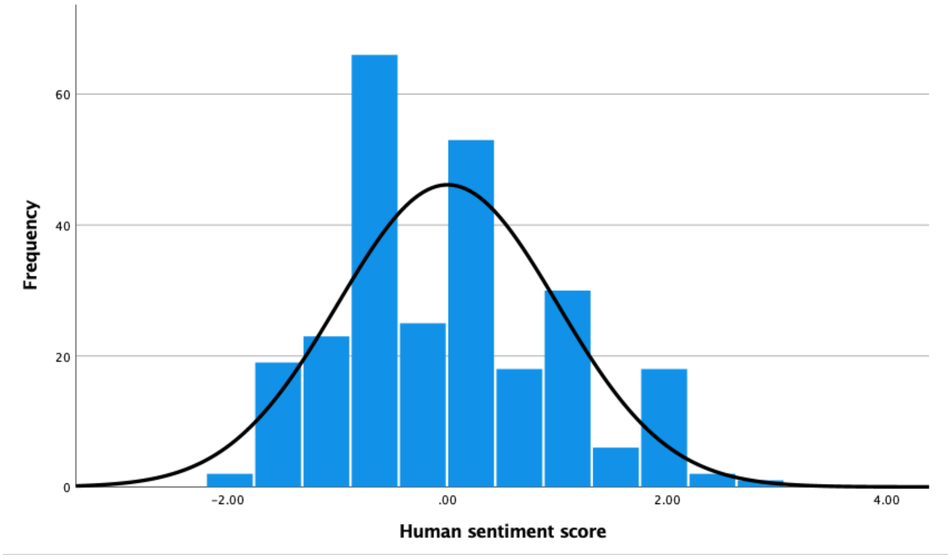
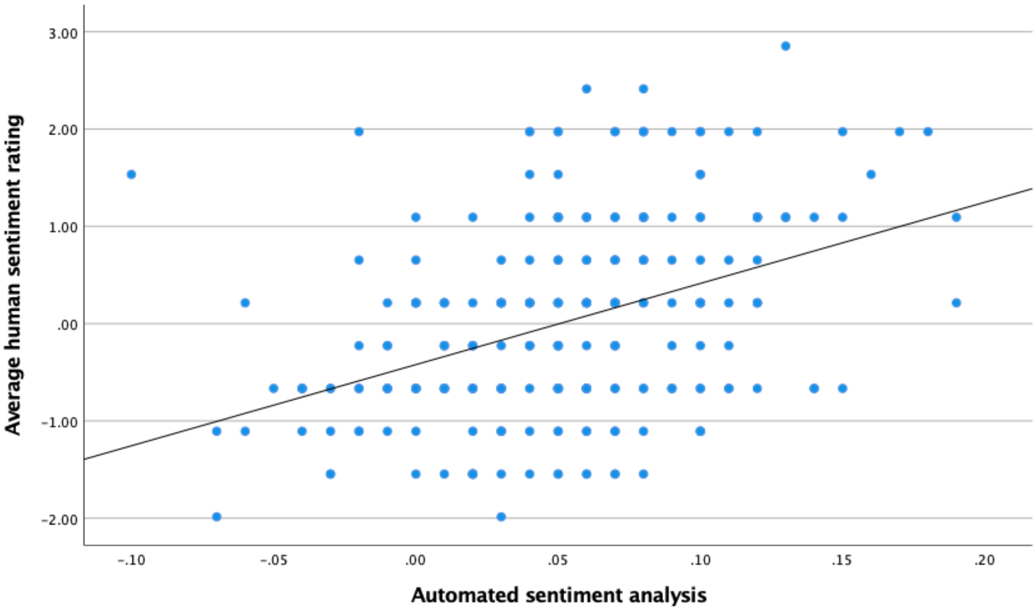


Figure 3

Scatterplot with linear fit line of the automated analysis (n = 315) and average human ratings (n = 263) regarding overall sentiment on the patient session records



Categorical comparison between the human raters and automated sentiment analysis

From Table 2 it is observed that the automated sentiment analysis rated a greater amount of session records as positive and less records as negative or neutral in comparison to the human raters regarding raw and standardised categorised sentiment scores (see Table 2). However, the standardised automated analysis' categorical scores do not differ considerably with the human raters on the categories 'positive' and 'negative' (see Table 2). Further, the human raters show similar ratings for each category with the largest difference for the category 'positive'. Lastly, from Table 3 it is observed that that the human raters had more true negative sentiment scores than true positives.

Table 2

Comparison of the raw categorical sentiment evaluations on the session patient records from the human raters and raw and standardised categorical sentiment evaluations from the automated sentiment analysis

	Rater 1 *N* (%)	**Rater 2 *N* (%)**	**Raw Automated Analysis *N* (%)**	**Standardised Automated Analysis *N* (%)**
Negative (%)	126 (47.0%)	127 (48.3%)	34 (10.8%)	116 (36.8%)
Neutral (%)	64 (23.9%)	70 (26.6%)	44 (9.6%)	64 (20.3%)
Positive (%)	78 (29.1%)	66 (25.1%)	237(75.2%)	135(42.9%)
Total	268	263	315	315

Table 3

Comparison between the human raters' categorical sentiment evaluations on the patient session records

		Rater 2			
		Negative	Neutral	Positive	Total
		N (%)	N (%)	N (%)	N (%)
Rater 1	Negative	106 (83.5%)	14 (20.0%)	5 (7.6%)	125 (47.5%)
	Neutral	14 (11.0%)	43 (61.4%)	4 (6.1%)	61 (23.2%)
	positive	7 (5.5%)	13 (18.6%)	57 (86.4)	77 (29.3%)
	Total	127 (100.0%)	70 127 (100.0%)	66 127 (100.0%)	263 127 (100.0%)

Automated-human agreement

Continuous Interpretation

The ICC(2,k) analysis revealed a moderate IRR (ICC = .55, CI = .43 – .65, F (262, 262) = 2.24, p < .001) between the automated analysis and average human ratings regarding overall sentiment on the session patient records (see Table 3).

Categorical Interpretation

The Weighted Cohen's kappa indicated a fair agreement, k = .293 (95% CI, .199 to .387, p < 0.001) for the agreement between the automated sentiment analysis and first human raters regarding overall sentiment on the session records.

Human-Human agreement

Categorical Interpretation

The Weighted Cohen's kappa indicated a substantial agreement, k = .68 (95% CI, .62 to .75), p = .000 between the human raters regarding overall sentiment on the session records.

Automated-human agreement per patient

Continuous Interpretation

The ICC (3,1) revealed a poor IRR for participant one diagnosed with OFSED and participant four diagnosed with AN, the ICC(3,1) was moderate for the remaining participants, including large confidence intervals(see Table 4).

Table 4

Intra-class correlation value for the agreement between the first human rater and the automated sentiment analysis for each participant

	Intraclass correlation	**95% Confidence interval**		**F-Test with true value 0**			
		Lower	bound	Value	df1	df2	significance
		Upper	bound				
Participant 1 (OFSED)	.24	-2.99	.55	1.31	55	55	.16
Participant 2 (AN)	.67	.43	.82	3.07	49	49	.00
Participant 3 (BN)	.63	-.60	.87	2.70	15	15	.03
Participant 4 (AN)	.42	-.09	.69	1.71	41	41	.05
Participant 5 (BED)	.53	.15	.75	2.14	43	43	.01
Participant 6 (BN)	.61	-.01	.85	2.58	18	18	.26
Participant 7 (BN)	.70	.18	.89	3.28	17	17	.01
Participant 8 (OFSED)	.60	-.08	.85	2.47	17	17	.04

Qualitative differences between the automated and human sentiment analysis

The sentiment scores of the automated and human analysis per patient over time can be found within Figure 4 to 11. Figure 4 shows a large difference between the average human rating and the automated analysis on session record 106, where the automated analysis showed a sentiment score of 4.0, but the human rater identified this record as irrelevant. Likewise, within Figure 5 the automated sentiment analysis showed a peak at record 209 whereas the human rater considered this record as irrelevant. The case of the automated sentiment analysis presenting a considerable larger sentiment score than the human rater is (almost) always coupled with the human rater appraising the session record as irrelevant.

Sentiment words specific to ED context. The automated sentiment analysis did not consider words specific to an ED context. An example can be seen within Figure 7 on session record 292 from a participant with AN where the human raters showed a sentiment score of 1.97 and the automated sentiment analysis a score of .40. When examining the positive and negative matches from the automated analysis regarding the session record, it was observed that the automated analysis did not rate certain context-specific positive ED words or expressions. For instance, the automated analysis did not rate the expression ‘beautiful recovery line’, ‘feeling more’, ‘taking space’. Again, the aforementioned example is not the only one encountered when examining differences between the human raters and the automated sentiment analysis. Therefore, a list with context-specific ED words and the different diagnosis can be found within Appendix D.

Lastly, the automated sentiment analysis categorised certain words to have a positive or negative polarity which were not considered or considered of the opposite sentiment within the human analysis. For example, the automated analysis indicated ‘exercising’ or ‘compensating’ as a positive match within the context of an AN diagnosis when, in fact, these expressions are mostly not of a positive polarity within a treatment context for AN. Moreover, the words ‘emotion regulation’ and ‘body experience’ were categorised as of a negative polarity which were not considered as sentiment-bearing words within the human analysis. Further differences regarding the positive and negative matches between the automated analysis and human raters can be found in Appendix E.

Figure 4

Sentiment scores from the automated sentiment analysis and the human sentiment analysis over time for participant one (OFSED) (N=175)

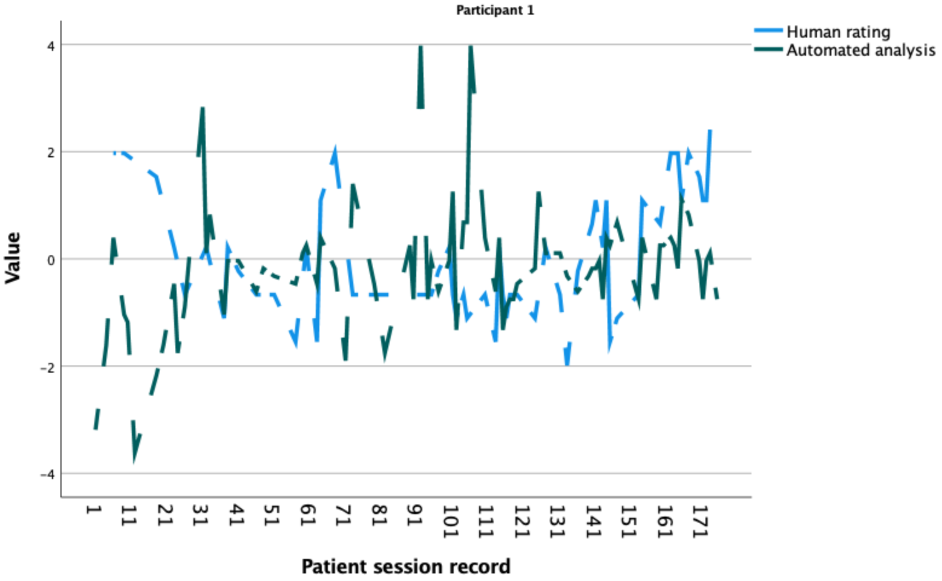


Figure 5

Sentiment scores from the automated sentiment analysis and the human sentiment analysis over time for participant two (AN) (N=68)

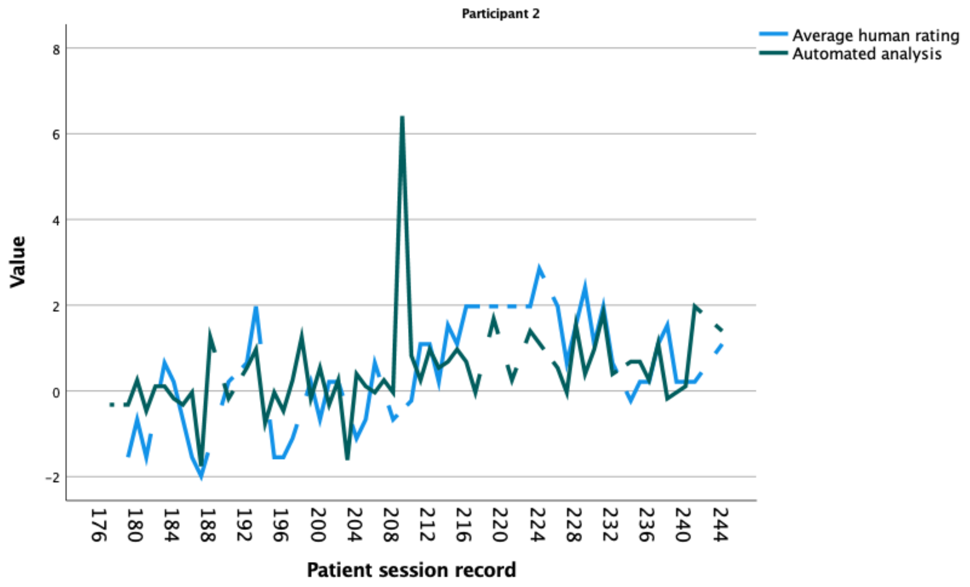


Figure 6

Sentiment scores from the automated sentiment analysis and the human sentiment analysis over time for participant three (BN) (N=22)

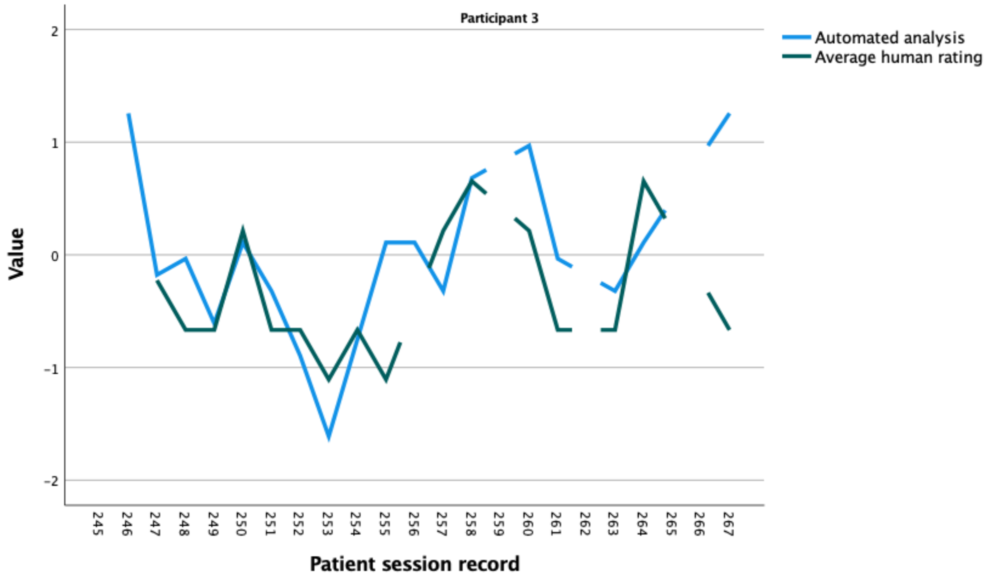


Figure 7

Sentiment scores from the automated sentiment analysis and the human sentiment analysis over time for participant four (AN) (N=61)

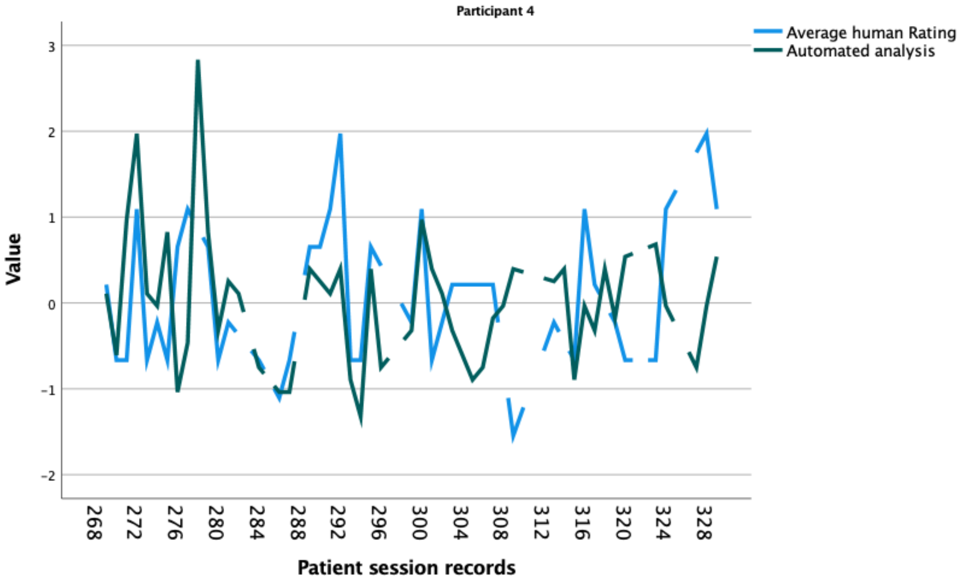


Figure 8

Sentiment scores from the automated sentiment analysis and the human sentiment analysis over time for participant five (BED) (N=56)

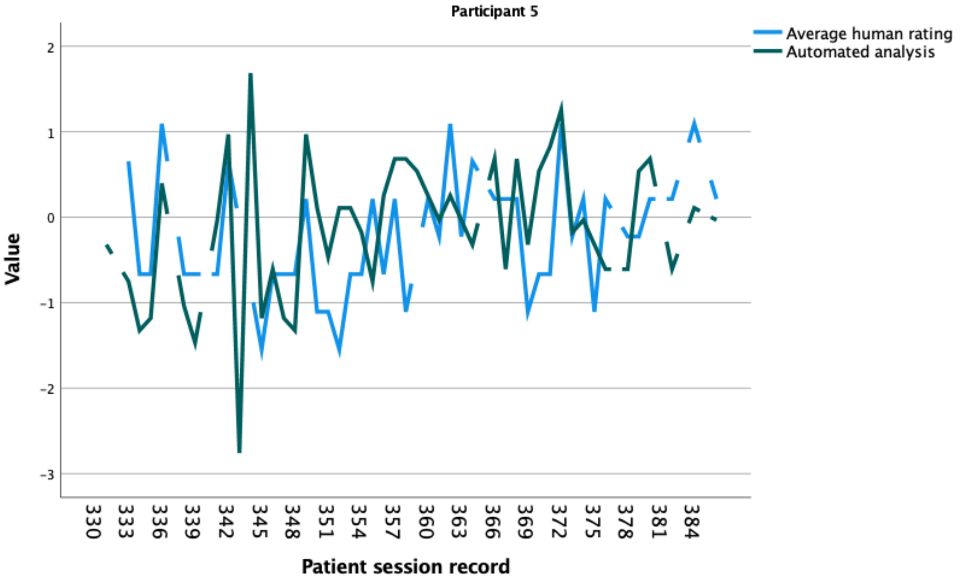


Figure 9

Sentiment scores from the automated sentiment analysis and the human sentiment analysis over time for participant six (BN) (N=30)

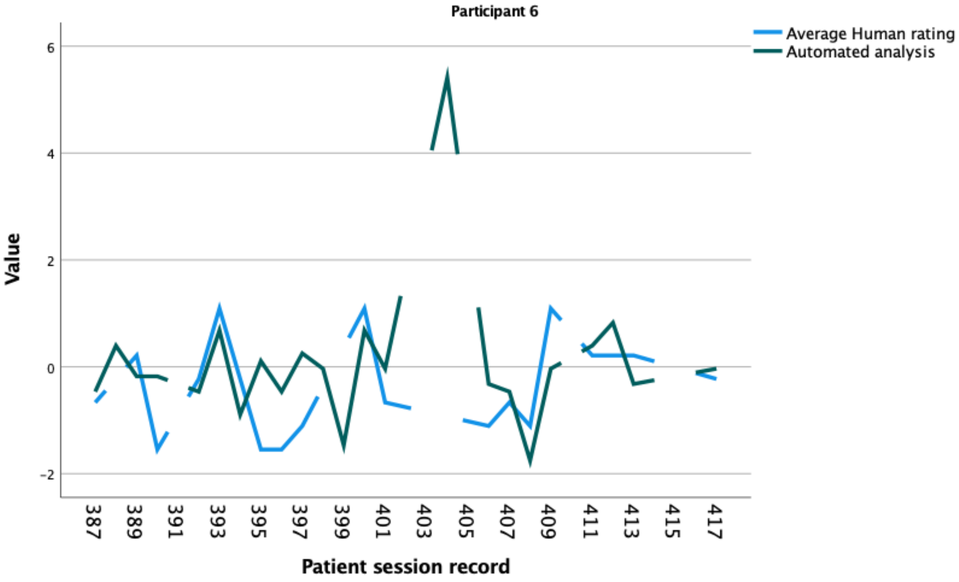


Figure 10

Sentiment scores from the automated sentiment analysis and the human sentiment analysis over time for participant seven (BN) (N=22)

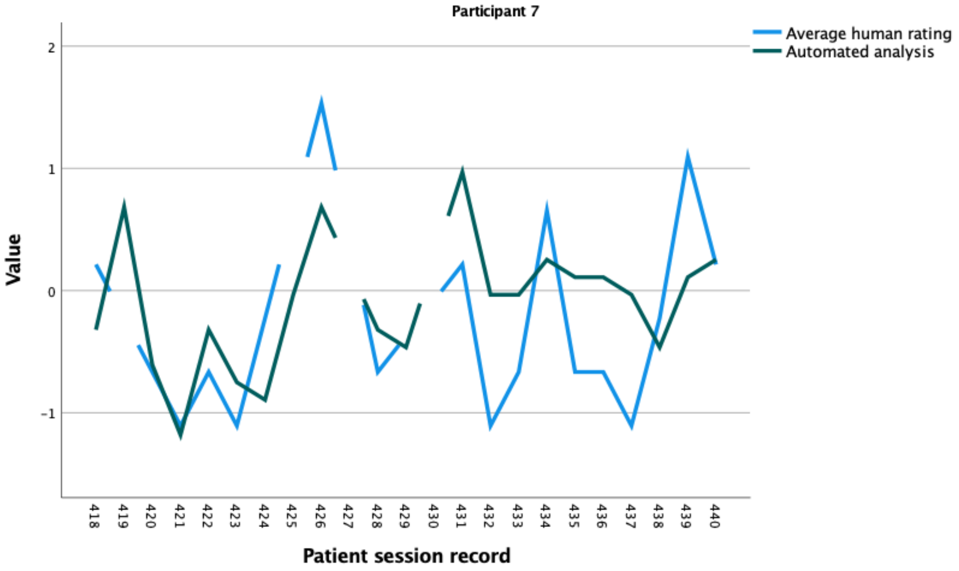
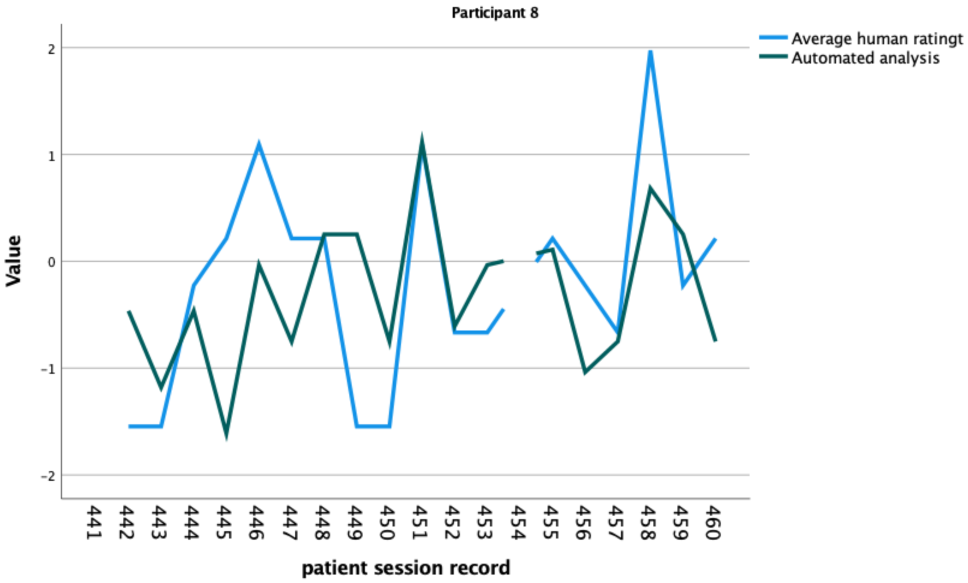


Figure 11

Sentiment scores from the automated sentiment analysis and the human sentiment analysis over time for participant eight (OFSED) (N=19)



Discussion

The aim of this study was to examine the performance of an automated sentiment analysis at extracting sentiment from session patient records within an ED treatment context compared to human raters. Additionally, the purpose of this study was to provide feedback to the designers of the automated sentiment analysis (6G) to optimise the analysis' future utilization potential. The results showed a moderate automated-human agreement under continuous interpretation (ICC = .55) and a fair agreement under categorical interpretation (k = .29) regarding the extraction of overall sentiment from the session records. The human-human agreement regarding overall sentiment was substantial under the categorical interpretation (k = .68). Further, the automated analysis scored the sentiment of the session records more positive than the human raters.

Automated-human agreement

The findings of the automated-human agreement are partly in line with other literature. While this study found a moderate continuous automated-human agreement and a fair categorical agreement, the exemplary study by Provoost et al. (2019) found a moderate automated-human agreement under both continuous and categorical interpretations. Further, a study investigating the performance of four different sentiment analyses compared to a baseline of multiple human raters found a fair automated-human agreement for three sentiment analyses and one with a moderate agreement, all under a categorical interpretation (Georgiou et al., 2015). Similarly, a study by Oksanen et al. (2015) found a fair automated-human categorical agreement between an automated sentiment analysis and each of three human raters all rating the sentiment of videos and comments related to AN.

The findings of the automated-human agreement could be explained by some shortcomings of the automated analysis. The automated analysis' lexicon did not include sentiment words specific to an ED context and may have indicated words that are of a negative nature as positive and vice versa. Moreover, the automated analysis assigned a sentiment score to words which were in itself not sentiment-bearing and therefore not considered by the human raters. Further, the automated analysis used 'n-grams' which only considered words before a sentiment-bearing word and not after and so, it may have overlooked the context of certain words. These shortcomings could have led to a diverging sentiment score and more positive rating of the records' sentiment compared to the human raters and hence, could explain the fair categorical agreement.

Further, other possible explanations may be due to the characteristics of the session patient records. The records included occasional misspellings or incorrect sentences, implicit statements of sentiment or varied in their length, content, and written language due to different clinicians. This will make the extraction of sentiment from the records by the automated analysis more complex and misinterpretation more likely, whereas human raters possess the ability and intelligence to comprehend difficult and ambiguous sentences and extract sentiment from these (Pang & Lee, 2008; Spinczyk et al., 2018). In addition, the session records mostly contained a summary of patients' difficulties and successes from the past days or weeks in between therapy sessions. As a result, the automated analysis' sentiment scores mostly centred around zero whereas most of the human raters' scores centred around light positive or negative. Moreover, seventy percent of the records classified as 'neutral' within the human analysis were also categorised as 'mixed', showing that sentiment may be difficult to extract from session records, often containing sentiment from both polarities. Furthermore, the sentiment within the session records is not directly stemming from the patients but is a clinician's interpretation of patients' sentiment and may therefore contain the subjective view of clinicians. While the automated analysis is not able to distinguish between patient's sentiment or clinician's view, human raters can, which could have resulted in the observed difference in sentiment ratings.

In summary, the automated analysis performed worse in discerning sentiment from session patient records in this study compared to the human raters. This means that the automated sentiment analysis cannot be used within practice.

Agreement between human raters

The finding of the substantial categorical human-human agreement is in line with other research which investigated the performance of an automated sentiment analysis against two or three human raters and found a substantial agreement as well (Moreno-Ortiz et al., 2019; Mukhtar et al., 2017; Wilson et al., 2005). In contrast, a moderate categorical human-human agreement was found within the study of Provoost and colleagues (2019) who used an average of eight human raters per texts.

A possible explanation for the findings could be due to the utilisation of a feedback session and clear protocol by the human raters. Likewise, a study by Moreno-Ortiz et al. (2019) incorporated a feedback session to optimise the used protocol and concluded a significant increase in the human-human agreement between the first and second trial, ensuring that the session records were rated in a similar manner. Further, both raters of this study possessed knowledge of EDs as they were both educated within the field of psychology. Hence, they may have understood words or expressions specific to an ED context and whether these were of positive or negative sentiment.

The human-human agreement within this study was chosen as a 'golden standard' to compare the performance of the automated analysis to as was the case within the aforementioned literature. Nevertheless, no perfect agreement has been found within literature regarding human-human agreement, meaning that human raters still lack consensus regarding the rating of texts' sentiment (Wilson et al., 2005). For this reason, it cannot be determined with certainty whether the automated analysis performed either 'good' or 'bad' as there is no solid benchmark.

Qualitative differences between the automated sentiment analysis and human raters

The automated sentiment analysis was tailored to the Dutch language and mental healthcare context but not to the context of EDs which resulted in a deviation in positive and negative matches between the automated analysis and the human raters. (see Appendices D and E). Further, the automated-human agreement of each patient showed a lower agreement for three participants with the diagnosis OFSED, AN and BED in comparison to the overall automated-human agreement. This means that the automated analysis did not encounter more difficulties with rating the sentiment within the context of a certain disorder.

Strengths and limitations

A strength of this study is that it is the first to examine the performance of an automated sentiment analysis on session patient records within a Dutch ED context. Further, the session records were written by trained clinicians concerning data from patients from an actual ED treatment centre, providing real contextual data. In addition, during this study, close contact was maintained with 6G upon which the automated analysis could be updated before the actual analysis. The findings of this study will also be provided as feedback to 6G to improve the automated analysis' performance. Furthermore, the utilisation of a feedback session may have supported that the records were rated in a similar manner by the human raters.

A limitation of this study was that it used less texts than other research investigating the performance of an automated sentiment analysis, as more than forty percent of the records within this study were not suitable for the analysis, decreasing the reliability of the results and possibly leading to a selective sample of records (Charter, 1999; Oksanen et al., 2015; Provoost, 2019; Wilson et al., 2005). Another limitation is that the human raters may be subjected to emotional bias which is a distortion in one's cognitions due to emotional factors such as personal feelings at the time of decision making (Yuan et al., 2019). Consequently, the affective state of the human raters at the time of rating the session records could have influenced the sentiment score that was given to a certain text. Further, this study only included two human raters which makes for a less representative interpretation of the overall sentiment within the session records in comparison to multiple raters (Stappen et al., 2021).

Future research and implications

Based on the findings it becomes apparent that the automated sentiment analysis cannot be used within practice to analyse the sentiment of session patient records if the gold standard of the average human ratings is considered an adequate indicator of the performance of the automated analysis. Ultimately, the aim is to use the automated sentiment analysis within practice to analyse the sentiment of individual patient's session records over time. However, the current study only showed moderate and fair automated-human reliability. No clinically relevant ICC value could be identified in the literature and therefore, although excellent reliability should be strived for, it is of interest to investigate at what ICC value automated-human reliability is good enough for the automated sentiment analysis to be used within clinical practice.

For future research it is advised to increase the number of human raters and examine the differences between the raters' sentiment scores in closer detail to improve the golden standard. Moreover, due to limited evidence regarding the validity of the utilisation of human raters as the gold standard, patients' own rating of their mood after or before therapy sessions or utilisation of patients' diaries and accompanying mood ratings could make a more solid benchmark to validate the automated sentiment analysis. Another key recommendation is to update the automated analysis' lexicon with context specific ED words and investigate its performance again on texts or session records within an ED treatment setting. Further, it is advised to manually remove irrelevant session records before the execution of the automated analysis since it is not able to determine the relevance of the records which could distort the overall sentiment score.

Further, the usability of patient session records for the extraction of patients' sentiment can be questioned due to its characteristics and lack of direct relation to the patient's sentiment. The sentiment of the session records and whether these could give an accurate representation of the patients' sentiment should therefore be investigated. However, despite the session records including complex and ambiguous information making them difficult to analyse, they do include valuable information about processes and underlying patterns contributing to EDs. Hence, it may be particularly interesting to use an open coding, through which the session records are examined on recurring ED themes which may be fruitful for the understanding of the mechanisms exhibited within EDs.

Conclusion

To conclude, the present study showed that an existing automated sentiment analysis performed worse than human raters in discerning sentiment from session patient records within a Dutch ED treatment context, therefore, the analysis cannot be used within practice. However, the human raters' also lack consensus on the evaluation of sentiment on the session records. This suggests that they may not be able to provide a solid benchmark to compare the automated analysis to, meaning another gold standard should be established such as patients' own mood rating. Moreover, the automated sentiment analysis requires to be optimised by including ED context-specific words within its lexicon to increase the accuracy of the analysis and needs to be further investigated. Lastly, it remains uncertain whether the session records are suitable for the extraction of patients' sentiment due to their complex and ambiguous nature, and it being an interpretation of the patient's sentiment by the clinician.

References

6Gorillas (2021). Het innovatieve dataplatform voor de zorg. https://6gorillas.nl/

Anderson, L. K., Reilly, E. E., Berner, L., Wierenga, C. E., Jones, M. D., Brown, T. A., Walter, H. K., & Cusack, A. (2017). Treating eating disorders at higher levels of care: Overview and challenges. Current Psychiatry Reports, 19(8), 1-9. https://doi.org/10.1007/s11920-017-0796-4

American Psychiatric Association (2013). Feeding and Eating Disorders. In *Diagnostic and Statistical Manual Of Mental Disorders (5th ed.)*. https://doi.org/10.1176/appi.books.9780890425596

Arcelus J, Mitchell AJ, Wales J, Nielsen S. Mortality Rates in Patients with Anorexia Nervosa and Other Eating Disorders: A Meta-analysis of 36 Studies. *Arch Gen Psychiatry, 68*(7), 724–731. https://doi.org/10.1001/archgenpsychiatry.2011.74

Basit, T. (2003). Manual or electronic? The role of coding in qualitative data analysis. *Educational research*, *45*(2), 143-154. https://doi.org/10.1080/0013188032000133548

Bagaric, M., Touyz, S., Heriseanu, A., Conti, J., & Hay, P. (2020). Are bulimia nervosa and binge eating disorder increasing? Results of a population-based study of lifetime prevalence and lifetime prevalence by age in South Australia. *European Eating Disorders Review*, *28*(3), 260-268. https://doi.org/10.1002/erv.2726

Ben-Zeev, D. (2017). Technology in mental health: creating new knowledge and inventing the future of services. *Psychiatric Services*, *68*(2), 107-108. https://doi.org/10.1176/appi.ps.201600520

Berends, T., Boonstra, N., van Elburg, A. (2018). Relapse in anorexia nervosa: A systematic review and meta-analysis. *Current Opinion in Psychiatry 31*(6), 445-455. https://doi.org/10.1097/YCO.0000000000000453

Berndt, D. J., McCart, J. A., Finch, D. K., & Luther, S. L. (2015). A case study of data quality in text mining clinical progress notes. *ACM Transactions on Management Information Systems*, *6*(1), 1-21. https://doi.org/10.1145/2669368

Boswell, J. F., Kraus, D. R., Miller, S. D., & Lambert, M. J. (2015). Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. *Psychotherapy research*, *25*(1), 6-19. https://doi.org/10.1080/10503307.2013.817696

Boulton, D. & Hammersly, M. (2006). Analysis of unstructured data. In R. Sapsford & V. Jupp (Eds.), *Data Collection and Analysis* (2nd ed., pp. 243-266). Sage Publications.

Charter, R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology*, *21*(4), 559-566. https://doi.org/10.1076/jcen.21.4.559.889

Chowdhary K. R. (2020) Natural Language Processing. *Fundamentals of artificial intelligence*. Springer. https://doi.org/10.1007/978-81-322-3972-7_19

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Dadvar, M., Hauff, C., & de Jong, F. (2011). Scope of negation detection in sentiment analysis. *Proceedings of the Dutch-Belgian Information Retrieval Workshop,* 16-20.

Daemen, D. M. (2020). De groep in tijden van corona*. Tijdschrift voor groepsdynamica & groepspsychotherapie*, *15*(4), 4-12.

de Beurs, E., den Hollander-Gijsman, M. E., van Rood, Y. R., Van der Wee, N. J., Giltay, E. J., van Noorden, M. S., van der Lem R., van Fenema, E.,. & Zitman, F. G. (2011).

Routine outcome monitoring in the Netherlands: Practical experiences with a web-based strategy for the assessment of treatment outcome in clinical practice. *Clinical psychology & psychotherapy*, *18*(1), 1-12. https://doi.org/10.1002/cpp.696

Denham, B. E. (2016). Interrater agreement measures for nominal and ordinal data. *Categorical Statistics for Communication Research* (pp. 232-254). John Wiley & Sons Inc. https://doi.org/ 10.1002/9781119407201

Devitt, A., and Ahmad, K. (2007). Sentiment polarity identification in financial news: a cohesion-based approach. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics* (pp. 984–991). https://doi.org/10.1.1.143.7157

Elliott, R. (2012). Chapter 6: Qualitative methods for studying psychotherapy change processes. In Thompson, A. & Harper, D. (Eds). *Qualitative Research Methods in Mental Health and Psychotherapy: A Guide for Students and Practitioners, (pp. 69 – 111). Wiley-Blackwells. https://strathprints.strath.ac.uk/42662/1/Elliott_2012_Qualitative_CPR_chapter.pdf*

Farooq, U., Mansoor, H., Nongaillard, A., Ouzrout, Y., & Qadir, A. M. (2017). Negation Handling in Sentiment Analysis at Sentence Level. *Journal of Computers, 12*(5), 470-478. https://doi.org/10.17706/jcp.12.5.470-478

Galmiche, M., Déchelotte, P., Lambert, G., & Tavolacci, M. P. (2019). Prevalence of eating disorders over the 2000–2018 period: a systematic literature review. *The American journal of clinical nutrition, 109*(5), 1402-1413.https://doi.org/10.1093/ajcn/nqy342

Georgiou, D., MacFarlane, A., & Russell-Rose, T. (2015). Extracting sentiment from healthcare survey data: An evaluation of sentiment analysis tools. *Science and Information Conference,* 352-361. https://doi.org//10.1109/SAI.2015.7237168.

Gilbody, S. M., House, A. O., & Sheldon, T. A. (2003). Outcome measures and needs assessment tools for schizophrenia and related disorders. *The Cochrane database of systematic reviews*, (1), 1-14. https://doi.org/10.1002/14651858.CD003081

Georgiou, D., MacFarlane, A., & Russell-Rose, T. (2015). Extracting sentiment from healthcare survey data: An evaluation of sentiment analysis tools. *Science and Information Conference* (pp. 352-361). IEEE. https://doi.org/10.1109/SAI.2015.7237168.

Hoek, H. W. (2016). Review of the worldwide epidemiology of eating disorders. *Current Opinion in Psychiatry, 29*(6), 336-339. Https://doi.org/10.1097/yco.0000000000000282

Hoerbst, A., & Ammenwerth, E. (2010). Electronic health records. *Methods of information in medicine*, *49*(04), 320-336. https://doi.org/10.3414/ME10-01-0038

Human Concern (z.d.). Ambulante behandeling. HumanConcern.nl. Geraadpleegd van https://humanconcern.nl/ambulante-behandeling/

IBM Corp. Released 2021. IBM SPSS Statistics for Macintosh, Version 28.0. Armonk, NY: IBM Corp.

Iliev, R., Dehghani, M., & Sagi, E. (2015). Automated text analysis in psychology: Methods, applications, and future developments. *Language and Cognition, 7*(2), 265-290. https://doi.org/10.1017/langcog.2014.30

Karpen, S. C. (2018). The social psychology of biased self-assessment. *American Journal of Pharmaceutical Education, 82*(5), 441-448. https://doi.org/10.5688/ajpe6299

Keel, P. K., Brown, T. A., Holland, L. A., & Bodell, L. P. (2012). Empirical classification of eating disorders. *Annual Review of Clinical Psychology, 8*(1), 381-404. https://doi.org/10.1146/annurev-clinpsy-032511-143111

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155-163. https://doi.org/10.1016/j.jcm.2017.10.001

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Lange, R. T. (2011). Interrater reliability. In J. S. Kreutzer, J. DeLuca, & B. Caplan (Eds.), *Encyclopedia of Clinical Neuropsychology* (p. 1348). Springer. https://doi.org/10.1007/978-0-387-79948-3

Larson, R., & Csikszentmihalyi, M. (2014). The experience sampling method. In *Flow and the foundations of positive psychology* (pp. 21-34). Springer. https://doi.org/10.1007/978-94-017-9088-8_2

Ledbetter, C. S., & Morgan, M. W. (2001). Toward best practice: leveraging the electronic patient record as a clinical data warehouse. *Journal of healthcare information management*, *15*(2), 119-132.

Lee, S., Xu, Y., D'Souza, A. G., Martin, E. A., D Doktorchik, C., Zhang, Z., & Quan, H. (2020). Unlocking the potential of electronic health records for health research. *International Journal of Population Data Science*, *5*(1), 1-9. https://doi.org/ 10.23889/ijpds.v5i1.1123

Maio, J. E. (2003). HIPAA and the special status of psychotherapy notes. *Professional Case Management*, *8*(1), 24-29.

Maluf, D. A., & Tran, P. B. (2008). Managing Unstructured Data with Structured Legacy Systems. Presented at IEEE Aerospace Conference, Big Sky, 1-8 March 2014. Piscataway: IEEE. https://doi.org/10.1109/AERO.2008.4526666

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*, *27*, 16-32. https://doi.org/10.1016/j.cosrev.2017.10.002

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, *5*(4), 1093-1113. https://doi.org/10.1016/j.asej.2014.04.011

Menger, V., Scheepers, F., van Wijk, L. M., Spruit, M. (2017). Deduce: A pattern matching for automatic de-identification of Dutch medical text. *Telematics and Informatics, 35*(4), 727-736. https://doi.org/10.1016/j.tele.2017.08.002

Moreno-Ortiz, A., Salles-Bernal, S., & Orrequia-Barea, A. (2019). Design and validation of annotation schemas for aspect-based sentiment analysis in the tourism sector. *Information Technology & Tourism*, *21*(4), 535-557. https://doi.org/10.1007/s40558-019-00155-0

Mukhtar, N., Khan, M. A., & Chiragh, N. (2017). Effective use of evaluation measures for the validation of best classifier in Urdu sentiment analysis. *Cognitive Computation*, *9*(4), 446-456. https://doi.org/10.1007/s12559-017-9481-5

Muzio, L. L., Russo, L. L., Massaccesi, C., Rapelli, G., Panzarella, V., di Fede, O., Kerr, A. R., & Campisi, G. (2007). Eating disorders: A threat for women's health. Oral manifestations in a comprehensive overview. *Minerva Stomatologica, 56*(5), 281-292. https://europepmc.org/article/med/17529915

Nikhil, R., Tikoo, N., Kurle, S., Pisupati, H. S., & Prasad, G. R. (2015). A survey on text mining and sentiment analysis for unstructured web data. *Journal of Emerging Technologies and Innovative Research*, *2*(4), 1292-1296.

Norman, S., Dean, S., Hansford, L., & Ford, T. (2014). Clinical practitioner's attitudes towards the use of Routine Outcome Monitoring within Child and Adolescent Mental Health Services: A qualitative study of two Child and Adolescent Mental Health Services. *Clinical Child Psychology and Psychiatry*, *19*(4), 576-595. https//doi.org/10.1177/1359104513492348

Oksanen, A., Garcia, D., Sirola, A., Näsi, M., Kaakinen, M., Keipi, T., & Räsänen, P. (2015). Pro-anorexia and anti-pro-anorexia videos on YouTube: Sentiment analysis of user responses. *Journal of medical Internet research*, *17*(11). https://doi.org/ 10.2196/jmir.5007

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundation and Trends and in Information Retrieval, 2*(1-2), 1-135. https://doi.org/ 10.1561/1500000001

Patel, V. L., Kushniruk, A. W., Yang, S., & Yale, J. F. (2000). Impact of a computer-based patient record system on data collection, knowledge organization, and reasoning. *Journal of the American Medical Informatics Association*, *7*(6), 569-585. https://doi.org/10.1136/jamia.2000.0070569

Percha, B. (2021). Modern clinical text mining: a guide and review. *Annual Review of Biomedical Data Science, 4*, 165-187. https://doi.org/10.1146/annurev-biodatasci-030421-030931

Provoost, S., Ruwaard, J., van Breda, W., Riper, H., & Bosse, T. (2019). Validating automated sentiment analysis of online cognitive behavioral therapy patient texts: An exploratory study. *Frontiers in Psychology, 10*, 1–12. https://doi.org/10.3389/fpsyg.2019.01065

Priyavrat, A. J. (2017). Sentiment Analysis: A Comparative Study of Supervised Machine Learning Algorithms Using Rapid miner. *International Journal for Research in Applied Science & Engineering Technology, 5*(6), 80 – 89.

Raja, U., Mitchell, T., Day, T., & Hardin, J. M. (2008). Text mining in healthcare. Applications and opportunities. *J Healthc Inf Manag*, *22*(3), 52-6.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Available at: https://www.r-project.org (accessed June 29, 2018). Royal HaskoningDHV versterkt data science-capaciteiten met de acquisitie van Ynformed. https://global.royalhaskoningdhv.com/nederland/projecten

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia, 126*(5), 1763-1768. https://doi.org/10.1213/ANE.0000000000002864

Schulte-van Maaren, Y. W., Carlier, I. V., Zitman, F. G., van Hemert, A. M., de Waal, M. W., van der Does, A. W., van Noorden, M. S., & Giltay, E. J. (2013). Reference values for major depression questionnaires: the Leiden Routine Outcome Monitoring Study. *Journal of Affective Disorders*, *149*(1-3), 342-349. https://doi.org/10.1016/j.jad.2013.02.009

Smink, W. A. C., Sools, A. M., Van Der Zwaan, J. M., Wiegersma, S., Veldkamp, B. P., & Westerhof, G. J. (2019). Towards text mining therapeutic change: A systematic review of text-based methods for Therapeutic Change Process Research. *Plos One, 14*(12), 1–21. https://doi.org/10.1371/journal.pone.022570

Spinczyk, D., Nabrdalik, K., & Rojewska, K. (2018). Computer aided sentiment analysis of anorexia nervosa patients' vocabulary. *Biomedical engineering online*, *17*(1), 1-11. https://doi.org/10.1186/s12938-018-0451-2

Stappen, L., Schumann, L., Sertolli, B., Baird, A., Weigell, B., Cambria, E., & Schuller, B. W. (2021). Muse-toolbox: The multimodal sentiment analysis continuous annotation fusion and discrete class transformation toolbox. In Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge (pp. 75-82). Association for Computing Machinery. https://doi.org/10.1145/3475957.3484451

Swinkels, I. C. S., Van den Ende, C. H. M., De Bakker, D., Van der Wees, Ph J., Hart, D. L, Deutscher, D., Van Den Bosch, W. J. H., & Dekker. J. (2007). Clinical databases in physical therapy. *Physiotherapy Theory and Practice 23*(3), 153-167. https://doi.org/10.1080/09593980701209097

von Holle, A., Poyastro Pinheiro, A., Thornton, L. M., Klump, K. L., Berrettini, W. H., Brandt, H., Crawford, S., Crow, S., Fichter, M. M, Halmi, K. A., Johnspn, C., Kaplan, A. S., Keel, P., La Via, M., Mitchell, J., Strober, M., Woodside, B. D., Kaye, W. H., & Bulik, C. M. (2008). Temporal patterns of recovery across eating disorder subtypes. *Australian & New Zealand Journal of Psychiatry*, *42*(2), 108-117. https://doi.org/10.1080/00048670701787610

Walfish, S., McAlister, B., O'Donnell, P., & Lambert, M. J. (2012). An Investigation of Self-Assessment Bias in Mental Health Providers. *Psychological Reports*, *110*(2), 639–644. https://doi.org/10.2466/02.07.17.PR0.110.2.639-644

Wampold, B. E. (2015). Routine outcome monitoring: Coming of age—With the usual developmental challenges. *Psychotherapy, 52*(4), 458-462. https://doi.org/10.1037/pst0000037

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proceedings of human language technology conference and conference on empirical methods in natural language processing, 375 – 354. https://aclanthology.org/H05-1044.pdf

Youn, S.J., Kraus, D. R., & Castonguay, L.G. (2012). The Treatment Outcome Package: Facilitating practice and clinically relevant research. *Psychotherapy, 49*, 115-122. https://doi.org/10.1037/a0027932

Yuan, J., Tian, Y., Huang, X., Fan, H., & Wei, X. (2019). Emotional bias varies with stimulus type, arousal and task setting: Meta-analytic evidences. *Neuroscience & Biobehavioral Reviews*, *107*, 461-472. https://doi.org/10.1016/j.neubiorev.2019.09.035

Appendix A

Consent form

INFORMEDED CONSENT WETENSCHAPPELIJK ONDERZOEK

ONDERZOEK: HET TOEPASSEN VAN TEKST DATA ANALYSE OM INZICHT TE KRIJGEN IN DE BEHANDELVOORTGANG BIJ EETSTOORNISSEN

Toelichting

Lees dit formulier alsjeblieft zorgvuldig. Als je anoniem gegevens van je behandeling beschikbaar wil stellen voor deze studie, dan kun je dat in dit formulier aangeven. De bijbehorende informatie brochure van het onderzoek kun je via deze link vinden. Hier kun je ook de contact gegevens van de onderzoekers vinden, mocht je vragen hebben over deelname aan deze studie.

Toestemmingsverklaring 1

Met de ondertekening van dit document geef je aan dat je minstens 17 jaar oud bent; dat je goed bent geïnformeerd over het onderzoek, de manier waarop de onderzoeksgegevens worden verzameld, gebruikt en behandeld en welke eventuele risico's je zou kunnen lopen door te participeren in dit onderzoek.

1. Ik kreeg voldoende informatie over dit onderzoeksproject. Het doel van mijn deelname in dit project is voor mij helder uitgelegd en ik weet wat dit voor mij betekent.

2. Mijn deelname in dit project is vrijwillig. Er is geen expliciete of impliciete dwang voor mij om aan dit onderzoek deel te nemen.

3. Mijn deelname houdt in dat gegevens die tijdens de behandeling verzameld worden gebruikt worden voor de doelen van het wetenschappelijke onderzoeksproject zoals beschreven in de informatiebrochure.

4. Het is mij duidelijk dat er bijzondere persoonsgegevens over mijn gezondheid en functioneren verwerkt worden en dat deze gegevens geanonimiseerd worden door de hoofdonderzoeker voordat er data-analyses plaatsvinden.

5. Het is mij duidelijk dat, als ik toch bezwaar heb met een of meer punten zoals hierboven benoemd, ik op elk moment mijn deelname, zonder opgaaf van reden, kan stoppen.

6. Ik heb van de hoofdonderzoeker de uitdrukkelijke garantie gekregen dat er voor wordt zorggedragen dat ik niet ben te identificeren in door het onderzoek naar buiten gebrachte

gegevens, rapporten of artikelen. Mijn privacy is gewaarborgd als deelnemer aan dit onderzoek.

7. Ik ben akkoord met eventuele (wetenschappelijke) publicaties die voortkomen uit dit onderzoeksproject en ben mij ervan bewust dat een er een wettelijke bewaartermijn van de anonieme data van 10 jaar geldt na publicatie.

8. Ik heb de garantie gekregen dat dit onderzoeksproject is beoordeeld en goedgekeurd door de Commissie Ethiek Psychologie van de Universiteit van Twente onder registratienummer ….

9. Ik heb dit formulier gelezen en begrepen. Al mijn vragen zijn naar mijn tevredenheid beantwoord en ik ben vrijwillig akkoord met deelname aan dit onderzoek.

10. Ik ben tenminste 17 jaar oud.

Ik ben **akkoord/niet akkoord** met bovenstaande punten.

Toestemmingsverklaring 2

Ik geef toestemming om mijn data, welke geanominiseerd wordt gearchiveerd als onderdeel van het beschreven onderzoek, beschikbaar te stellen voor toekomstig onderzoek en lesdoeleinden door anderen onderzoekers en docenten. Ik begrijp dat deze gegevens nooit te herleiden zijn tot mij als individu.

Ik ben **akkoord/niet akkoord**.

Appendix B

Information brochure

https://humanconcern.nl/wp-content/uploads/2022/05/Informatiebrochure-wetenschappelijk-onderzoek-2022.pdf

Appendix C

Protocol

Protocol for the human sentiment analysis

In order for the human raters to assess the texts in an objective and similar way, a protocol was created to analyse the patient session record texts to attribute a sentiment score to each. Therefore, the following points must be considered when analysing the texts:

1. Texts with less than five words will not be examined, nor will patient session records including information about work supervision, communication and arranged appointments with other clinicians or institutions and descriptions of actions taken by the clinician(s) regarding administrative activities.
2. Patient session record texts including phone calls, voicemails or mobile texts containing sentiment will be considered regarding sentiment.
3. *Results* written by the clinician about the patient's sentiment will be considered, as it contains sentiment from the patient.
4. When rating the patient session record, the diagnosis of the patient must be known and considered.
5. The examination and ratings of a patient session record text should be based, only, on the given patient session record text, previous texts belonging to the same patient and context outside a given text should not be considered when rating a certain text. However, metaphors and the meaning behind indirect sentiment will be considered.
6. The examinations and ratings of the patient session record texts should be based on the sentiment of the patient. Parts within the patient session record texts about the clinician's sentiment, subjective view, or treatment instructions (what the patient is going to do next) should not be considered.

Appendix D

Wordlist with words specific to the contexts of EDs

Table 6.

Positive and negative context-specific sentiment words regarding ED from the human analysis

	Sentiment	
	Positive	Negative
General	Uitdagingen aangaan Eet-uitdagingen Herstel Hulp vragen Regie nemen Open over emoties / emoties delen Gezonde kant / gezonde gedachten Verboden eetlijst proberen Deelnemen aan het leven Genieten (van eten) Dankbaar(heid) Minder regels Behoeftes uitspreken / grenzen aangeven / voor jezelf kiezen Gunnen	(Voor)compenseren / eten overslaan 'Eetstoornis trekt' 'Eetstoornis nodig hebben' 'Eetstoornis opspelen'/ last van eetstoornis Eetstoornis is heftig / aanwezig' 'Niet delen hoe het gaat'/ onderdrukken van emoties /emoties niet delen Braken/ braakgedachtes Eetstoornis gedachtes Schuldgevoel Innerlijke criticus Overspoeld door emoties Niet meer in de hand Vastzitten 'eetstoornis-stem'

	Eetstoornis onderdrukken /	Regels
	eetstoornis op de	Eetbui / overeten
	achtergrond	Overdenken
	Meer voelen	Afvallen
	Trots op zichzelf	Veilige keuze
	Angst loslaten	Zichzelf groot houden
	Kwartje gevallen	Dik voelen / angst dik
	Emotionele lading minder	worden
	Besef	Schaamte
	Luisteren hongergevoel	Escape
	Toegeven emoties	Onrust
	Grip hebben	Negatieve
	Aankomen	lichaamsbeleving
	Openheid	Weinig eten
	met zichzelf in conact	Niet luisteren naar grenzen
	dingen aangaan	Restrictief eten
	minder wandelen / stappen	Niet bewust van honger
	zetten	gevoel / trek
	bewustworden	Vreselijk om voor spiegel
	groei doormaken	te staan
	opgewekt	Kritisch op lijf
	vrijheid	Eetbui-drang
	rust	Lijdensdruk
	kracht om tegen eetstoornis	Terugval / vervallen oude
	in te gaan	patroon

	niet extra sporten Normaal eten geen paniek momenten emotieregulatie herstellijn meedoen met anderen flexibel zijn gewich doet de patient niks	Overleven Uiterlijk controleren / controledrang Afwezig contact niet gezien voelen prestatiedrang bewijzen aan zichzelf wandelen / stappen obsessief
AN	Afbouwen sporten Aankomen	Sporten (en elke vorm die daarbij komt kijken)
BN		Braken / braak gedachtes / overgeven Eetbui / overeten
BED	Sporten Afvallen	Eetbui / overeten
OSFED	Afbouw sporten	

Appendix E

Sentiment words categorized by the automated sentiment analysis as of positive or negative polarity but not within the human analysis

Table 7

Sentiment words categorised by the automated analysis as a positive or negative match which were not within the human analysis

Positive	negative
Sporten	Lichaamsbeleving
Compenseren	Emotieregulatie
Controle	Geur
Bewegen	Adhd
Bekend	Instelling
Waarneming	Te ervaren
Baan	Systeem
Reis	Geen therapie
Bekend	Klein
Soort	Te bereiden
Beleving	Vet
Buiten	Klinische
Vakantie	Kwetsbaar
Aanpassen	Brood
Kenmerken	Kistje
Definitief	
Informatie	
Rest	

Kind
Vorm