UTRECHT UNIVERSITY
Faculty of Science
Department of Information and Computing Sciences
MSc Artificial Intelligence

A LINGUISTIC APPROACH TO DETECTING WORKING ALLIANCE IN THERAPIST-PATIENT PSYCHOTHERAPY

A THESIS BY
Lennard R. Bornemann
9993029

Project supervisor Prof. dr. Albert Salah
Second supervisor Dr. Sanne J.E. Bruijniks
Second examiner Dr. Itir Önal Ertuğrul

Acknowledgements

I would like to first and foremost thank my family and friends for their continued support, without which this project would not have been possible. They not only helped to inspire me, but also to persevere in difficult times.

I would also like to thank my supervisor Professor Dr. Albert Salah for his guidance and support throughout this process, teaching me to look at problems with new perspectives.

Additionally, I extend my gratitude to Dr. Sanne Bruijniks, from whom the data used here originates, as she also provided useful insight and direction through her own research requirements.

Lastly, I would also like to thank my second reader Dr. Itir Önal Ertuğrul whose proposal feedback guided the start of this project.

Abstract

A central component determining the outcome of psychotherapy is the working alliance between a therapist and their patient. This study explores an approach to understanding and predicting this alliance through automatic linguistic analysis of therapy session transcripts, as opposed to traditional self-report measures such as the working alliance inventory.

The methodology encompasses a pipeline for the automatic transcription, diarization and identification of participants in therapy sessions, followed by affect and sentiment analysis at a speaker-level using a custom-trained Dutch language model, which shows promise in capturing affective trends.

Following this analysis, several features are extracted such as the emotional valence, arousal, sentiment and speaker synchronies, aiming to predict working alliance inventory scores and their sub-components of bond, goal and tasks. Significant correlations between some of these features and the alliance scores are revealed, particularly a patient's average valence.

While the final predictive power of the presented models is lacking, valuable insights are gained into the issues surrounding such automatic analysis and prediction. The contribution of this study to computational psychotherapy research is therefore mainly a proof of concept for language based working alliance evaluation. Relevant code is available at: `https://github.com/LrbstudyUU/UUMasterThesis`

Table of Contents

List of important Abbreviations

Abbreviation	Full Form
WAI	Working Alliance Inventory
WAI-S	Working Alliance Inventory-Short
WAI-SRT	Working Alliance Inventory-Short Revised Therapist
CBT	Cognitive Behavioral Therapy
IPT	Interpersonal Psychotherapy
NLP	Natural Language Processing
BERT	Bidirectional Encoder Representations from Transformers
LIWC	Linguistic Inquiry and Word Count
rLSM	Reciprocal Linguistic Style Matching
VAD	Voice Activity Detection
SCD	Speaker Change Detection
OSD	Overlapped Speech Detection
WER	Word Error Rate
DER	Diarization Error Rate
JER	Jaccard Error Rate
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error

1. Introduction

The concept of therapeutic alliance, or working alliance encompasses the professional relationship between a therapist and their patient through a focus on assigned goals, tasks and their shared bond, as proposed by Bordin (1979). This alliance is fundamental to the therapeutic process, as demonstrated by Horvath and Symonds (1991), who found that a strong alliance is indicative of successful therapeutic outcomes. The ability to gauge the strength and quality of this relationship offers therapists a valuable tool allowing them to better understand and improve on their therapeutic approach.

A widely used evaluation system for this working alliance is the working alliance inventory, which is a self-report questionnaire developed by Horvath and Greenberg (1989). By having this questionnaire filled out by the therapist, patient and an outside observer, the perception of the therapeutic relationship can be measured through three different perspectives. While the manual completion of such self-report surveys represents the current standard for evaluating the working alliance, this practice is time-consuming, which provides motivation for the development of systems that are able to automatically identify and predict such alliance.

This thesis ventures into creating such an automatic approach by identifying and extracting verbal linguistic features from video recordings of therapy sessions in an attempt to predict the working alliance between therapist and patients.

Given the sensitive nature of psychotherapy, the availability of such recordings is quite limited, leading to restrictions in the scope of the proposed automatic approach. While details of the data used in this thesis will be elaborated upon further in Chapter 3, it is important to mention that it is a relatively small Dutch dataset of recordings, meaning that the findings of this research are unlikely to apply across the globe.

It is also important to mention that while an automatic prediction of working alliance via machine learning may reduce the subjectivity surrounding the human perception and prediction of working alliance, it is not guaranteed that such automatic prediction is free from bias. This is because the data used to make predictions still originates from humans, who always carry some degree of bias based on their lived experiences, stereotypes and their current state. Prediction models therefore represent a less subjective average of these biases rather than being completely free of them.

One of the main intrinsic challenges the dataset used in this thesis carries is its video quality, as the presence of both therapist and patient is not guaranteed in the majority of available recordings, further supporting a text-based approach to predicting working alliance via transcription of these videos.

1.1 Research questions

The following question represents a starting point in predicting the working alliance in psychotherapy:

Research question 1: What linguistic features are relevant to the therapeutic working alliance?

The nature of this question is multifaceted as relevance can stem from professional opinion as well as computational outcomes, although here it is taken to mean features that aid the prediction of the working alliance. Chapter 2 covers related research which seeks to answer this question and provide a foundation of possible predictor features. Here it is also important to reiterate that the working alliance is captured using a popular evaluation survey called the working alliance inventory, the details of which are also covered in Chapter 2. This distinction is important, as although the working alliance inventory has been widely used and regarded as a strong descriptor of the working alliance, the features which carry relevance in predicting the items on the survey may not be fully equal to those which represent the true working alliance between therapist and patient. An important question to consider alongside the relevancy of features is the following:

Research question 2: How can linguistic features be extracted from video data?

This is an important consideration, as it also expands beyond the video format. In real life video and audio recordings are likely the easiest way to capture and represent conversations and therapy sessions. It is unlikely that such sessions are transcribed as they are happening and even less likely that any features such as the emotional state of a participant is recorded in real time. Therefore it is important to consider how relevant features can be extracted from recordings. While identifying and extracting possible predictor features is the basis of prediction, the following question arises:

Research question 3: To what extent can linguistic features be used to predict this working alliance?

This is because the simple presence of certain predictor features does not directly translate

into a perfect prediction of the associated working alliance. Therefore investigating the extent to which prediction is possible with these features is crucial as it would allow future research to focus on or disregard similar approaches.

2. Related Work

2.1 Working alliance

In the field of psychotherapy, the concept of a working alliance between a therapist and their patient has been established by Bordin (1979), in order to capture their therapeutic relationship as well as serving as a possible predictor of therapy outcome.

Bordin postulates that the working alliance is comprised of three key components; goals, tasks and the bond between therapist and patient. Goals here refer to shared and agreed upon desired outcomes for the therapy. Here Bordin gives an example about how a specific goal in behavioural therapy would be to change an individual's behaviour in connection to others or their environment. The alignment of goals can be assessed by both parties via discussion about their expectations.

In order to work towards achieving these goals, tasks are required, referring to specific therapeutic interventions and activities that are employed or assigned. Following the earlier example of behavioural therapy by Bordin, appropriate tasks would encompass bringing about the desired behaviour change through positive reinforcement provided by the therapist or others.

Finally, the bond component of the working alliance concerns the professional relationship between therapist and patient. While Bordin notes that a basic level of trust is required for any therapeutic relationship, delving into deeper recesses of an individual's experiences is facilitated by a respectively deeper level of such trust and attachment to the therapeutic counterpart is required. All three elements can be evaluated via ongoing discussions between the parties involved in therapy, as well as through self-report measures such as the working alliance inventory which will be discussed in further detail in a later section.

Overall, the working alliance was seen by Bordin as a key to the change process in psychotherapy, maybe even the most important one. Since then, there have been multiple studies which support this idea. A meta-analysis performed by Horvath and Symonds (1991) compared 24 studies based on 20 separate datasets to reach the conclusion that a reliable relationship between working alliance and positive therapy outcome seems to exist.

This relationship draws further support from yet another meta-analysis also conducted

by Horvath et al. (2011), in which the results of around 200 reports were compiled, leading to a similar conclusion about working alliance being an important factor playing into therapeutic outcome.

Adding yet further support, a look at the work of Castonguay et al. (2006) highlights important conclusions about working alliance that have been found in previous research, importantly the reiteration of the relation between working alliance and outcome of therapy, but also that poor early working alliance seems to be predictive of client dropout.

In conclusion, the working alliance is a central theme in psychotherapy research, with growing evidence supporting its seemingly critical role in facilitating positive therapy outcomes.

2.2 Working alliance inventory

Based on the concept of working alliance described above, Horvath and Greenberg (1989) developed a questionnaire, the Working Alliance Inventory (WAI), aimed at capturing this alliance in a comprehensive way. This survey consists of 36 questions spanning the perceived interpretations of topics such as the relationship of the patient and their therapist, tasks and goals of the therapy.

The items in the WAI are evaluated using a five-point Likert scale to reflect the level of agreement with each item, with options ranging from "seldom" to "always". An example item from the WAI is; "I feel that my therapist(s) appreciates me.", demonstrating how the bond between therapist and patient is expressed in a manner that can be answered using the mentioned Likert scale. Further examples can be seen in Appendix A.1 and A.2.

In order to measure the working alliance using the WAI, both patients and therapists need to complete the survey after a session. Additionally, an observer is also usually tasked to fill out the questionnaire in order to gain an objective perspective on the relationship.

It is important to note that a shorter version of the WAI exists, proposed by Tracey and Kokotovic (1989), which aims to perform the same task of capturing the working alliance in a more compact fashion. This version of the survey consists of only 12 questions, which were chosen by selecting four questions from the original WAI with the strongest correlations to each aspect of the working alliance that were measured; goals, tasks and bond. This version is also called the Working Alliance Inventory-Short (WAI-S), for which a similar shortened version exists for therapists, called the Working Alliance-Short Revised Therapist (WAI-SRT), introduced by Hatcher and Gillaspy (2006).

The validity of the interchangeability of the WAI and WAI-S has been demonstrated also by Busseri and Tyler (2003) in a meta-analysis spanning 54 university counselling centres. The findings indicate that the scores of the WAI and WAI-S were highly correlated and had similar descriptive statistics, pointing to the fact that using the WAI-S for time convenience purposes and ease of use is a reliable approach to measuring the working alliance in place of the complete WAI. This interchangeability lends itself to automatic working alliance assessment, as it translates to a reduction in complexity as fewer target variables need to be predicted.

An important consideration about the WAI scores is highlighted by Vollebregt (2023), in that the perception of working alliance differs for the three rating groups; patients, therapists and observers. This subjectivity, also noted by Bordin (1979) needs to be kept in mind, as different linguistic features likely have different correlations to each group's perception.

2.3 Relevant concepts

As demonstrated previously, the concept of working alliance is pivotal in the domain of psychotherapy, focusing on the goals, tasks and bond that exists between a therapist and their patient within a therapeutic environment. However, it is important to acknowledge that other constructs also play significant roles in determining therapeutic outcome.

One such element is trust, which is an essential part of any relationship. Although the working alliance present in therapy should not be directly equated to a personal relationship between therapist and patient, the element of trust still represents a vital foundation in its existence.

This notion is underscored by research conducted by Fuertes et al. (2017), in which the authors outline that trust is a core component of working alliance as it is important for patients to see their therapist as an "ally". The reverse is also true, as therapists should also see their patients as such to better establish and navigate common tasks and goals. This mutual perception of alliance also encourages the lowering of patient inhibition to sharing personal information and traumatic experiences, which is necessary for their healing process.

Trust is not merely a distant concept, but can be quantitatively assessed in its own way, through tools such as the Trust in Physician scale (TIPS) introduced by Anderson and Dedrick (1990). Despite being measurable on its own, the measurement of trust exhibits certain overlaps with the WAI. This is because trust can be built through the establishment

of well-defined goals and tasks.

Statements from the TIPS such as; "My doctor is a real expert in taking care of medical problems like mine." and from the WAI like; "I believe the way we work on my problems is the right one." demonstrate such overlap. Both statements draw on a level of trust that is expressed by the patient about their caretaker's ability to help.

The relationship between trust and working alliance can be better understood by delving into the concept of therapeutic rapport. Therapeutic rapport refers to the forging of a trusting and empathetic relationship between therapist and patient. The significance of rapport in establishing the above-described openness to share information through trust is underscored in the research of Newhill et al. (2003).

The main difference between the concepts of working alliance and therapeutic rapport lies in the fact that the former fixates on goals and tasks as well as a bond between therapist and patient as previously mentioned, whereas the latter mainly refers to the quality and depth of their relationship.

The nuanced interplay of these concepts is emphasised by Rogers (2015), in which the author expresses a correlation between rapport and working alliance, suggesting that the working alliance encompasses aspects of rapport. This indicates that a good therapeutic rapport can serve as the foundation for a strong working alliance.

2.4 Communication

Effective communication is a necessary part of providing health care in a way that leads to improved health outcomes. This principle is emphasised by Mauksch et al. (2008) in an investigation about communication in medical encounters. Although this investigation extends across various medical fields beyond psychotherapy, the concepts that are deemed important are largely similar. These elements include the development of rapport, strategic agenda-setting and an empathetic sensitivity to social and emotional cues.

Here it is important to emphasize that communication is not solely based on spoken word, but rather a combination of verbal and non-verbal factors. In an article by Berry and Pennebaker (1993), the authors highlight the fact that emotional communication occurs through both non-verbal factors such as a patient's gaze, their body posture or even their vocal qualities, while the verbal side of emotional communication tackles the translation of thoughts into words. An interesting aspect of the factors of communication denoted in this article is that both forms of expression lead to reductions in autonomic nervous system

activity. This suggests that both verbal and non-verbal communication facilitate the human body's ability to reduce stress and calm itself.

These insights add validity to the importance of vocal communication, especially as it does not only convey information to the therapist about the patient and their problems but also serves as a bridge of general communication on a level of importance as non-verbal factors, allowing for a more solid working alliance to be built.

The importance of verbal factors in communication, especially in building rapport is highlighted in a study conducted by Bronstein et al. (2012). While the focus of this study is general negotiation, the concept of rapport is largely the same as laid out previously, suggesting that the findings hold relevancy in the context of working alliance. The authors note that rapport can be expressed through some key verbal behaviours such as expressions of positivity, coordination and mutual attention. These behaviours can be identified through linguistic elements such as politeness, reflections of emotion and restatements, which refer to the repetition of certain words or phrases.

Here the research of Negri et al. (2019) is of interest, in which the authors suggest a correlation between higher levels of emotional language and positive emotional expression used by patients and a strong working alliance after their first therapy session. While the status of a therapy session can not always be known to be the first to outsiders such as researchers, such a correlation still suggests that high levels of emotional language carry importance for the working alliance.

A meta-analysis by Pinto et al. (2012) expands on this by evaluating the influence of various verbal and non-verbal communication factors on working alliance. Notably, verbal interactions characterized by warm greetings, attentiveness, clear explanations, gentleness, reassurance, and engaging questions were associated with stronger alliances, underscoring the importance of verbal communication in therapeutic success.

Another important component of communication is linguistic alignment, or synchrony, proposed by Brennan and Clark (1996). This refers to the phenomenon in which individuals adapt their language style, vocabulary and even speech patterns to mirror their conversational partner to some degree.

Further investigations conducted by Branigan et al. (2007) suggest that such alignment occurs regardless of conversational group size, although a sensitivity to participation role does seem to be present, meaning that synchrony is strongest between the primary conversational entities.

Also, linguistic alignment is often associated with reciprocal understanding and rapport building. This is underscored by the research of Duran et al. (2019), suggesting that it reflects the interpersonal relationship present in a dyad, which relates back to the concept of working alliance.

The link between linguistic alignment and working alliance is also underscored by the work of Bayerl et al. (2022). Here the authors analyse dyadic therapy sessions in an effort to better understand the effect of linguistic alignment on the working alliance in psychotherapy. Their results show that higher alignment correlates with higher WAI scores, suggesting that linguistic alignment is indeed a marker of the therapeutic alliance.

Similar findings were made by Vail (2023), as a section of their research also focuses on how such synchrony between therapist and patient influences the patient's perception of their working alliance. The outcome of this analysis supports the research of Bayerl et al. (2022), as the results suggest that the alignment of the patient can act as a strong predictor of the working alliance. Additionally, a high alignment of the therapist is also associated with a better perception of the working alliance from the patient's point of view, further cementing the importance of linguistic alignment in evaluating therapeutic alliance.

Furthermore, a study conducted by Sharma and De Choudhury (2018) delves into the linguistic behaviours of online mental health communities. Here the authors note that such communities develop into safe places for their members in part through linguistic alignment, as it promotes social cohesion, trust and empathy.

2.5 Affect

While analysing and extracting specific emotions represents one approach to learning more about the emotional content of text, another approach relies on the affective concepts of valence and arousal.

Valence refers to the positivity or negativity of an emotion, based on the circumplex model of emotions introduced by Russell (1980). In the context of text analysis, valence can help determine whether the language conveys happiness, sadness, anger and other emotions. For example, words like "happy," "love," and "excited" are associated with positive valence, while words like "sad," "hate," and "angry" are associated with negative valence.

This concept differs slightly from sentiment, as valence focuses more on the emotional expression, whereas sentiment describes the overall nature of a text. Since psychotherapy can deal with traumatic experiences, valence may capture the levels of different emotional

experiences, while the sentiment may be influenced by the negative context of the session. It is therefore important to analyse both of these aspects.

Arousal, on the other hand, measures the intensity of emotions. Again based on the emotional model of Russell (1980), arousal ranges from calm to excited or agitated, evaluating how much an emotion stimulates the individual. For example, words like "excited" or "frightened" indicate high arousal, while words like "calm" or "relaxed" suggest low arousal.

It is important to note that the example words provided above are inherently emotive to help illustrate the concepts of valence and arousal. However, these concepts are not limited to such emotive language. Rather, they represent a dimensional view of emotion. This dimensional view is captured by the previously mentioned model of Russell (1980), in which valence and arousal form the X and Y axes. Each word in a language is therefore seen to carry some distribution of these two core dimensions. An example of this is the affective norms for English words (ANEW) dataset by Bradley and Lang (1999), in which several thousand English words spanning beyond classic emotive examples are rated for their valence and arousal as well as dominance. Another example is a Dutch emotion, valence and arousal dataset by Speed and Brysbaert (2023), the details of which are covered in Chapter 3.

Dominance refers to a degree of control over a situation or person in an emotional context and is sometimes used as an additional dimension when capturing emotion. This dimension is however secondary in its importance in emotional variance when compared to valence and arousal, as noted by Bradley and Lang (1994), due to its relational nature. This means that dominance ratings can change drastically depending on its interpretation and are therefore less reliable than the primary dimensions. An example Bradley provides is that of a snake, which was rated to have high dominance by some participants due to its dangerous nature, while also being given low dominance scores by others who instead saw it to be the reason, as a human, to feel less in control and therefore invoke feelings of low dominance. Valence and arousal will therefore also remain as the primary dimensions investigated here.

2.6 Multimodal machine learning

For the context of this research, it is important to outline the existence and use of multimodal models. Unlike unimodal models, which consist of a single type of input to produce an output, multimodal models are able to combine multiple types of input. Such models, described by Baltrušaitis et al. (2018) are distinguished by their ability to process and integrate information from a variety of sources or modalities, such as text, visual, audio,

and sensory data. This integration is crucial, as it mirrors human information processing more closely than unimodal systems are able to, allowing for the development of more sophisticated and capable AI systems.

In the realm of therapeutic working alliance, such multimodal learning is utilised in the research of Vollebregt (2023), with the same data provided by Bruijniks et al. (2020), in which a combination of visual, textual and audio features are merged to predict the working alliance from the perspective of patients, therapists and observers. Unfortunately, none of the models used (multilinear regression, XGBoost, k nearest neighbours, random forest, support vector regression and Elastic net) were able to predict the alliance well, although this lack of performance was likely caused due to a lack of appropriate data since only a fraction of the dataset lent itself to the multimodal analysis preceding the predictions.

Regardless of this shortcoming, a key takeaway from Vollebregt (2023) is that certain textual features were among the features with the highest importance for the predictions. For patients, emotions such as amusement, approval, fear and curiosity were the top predictors as well as a minimal arousal from the therapist. Curiosity also played an important role in the prediction of therapist scores alongside minimal valence of the patient. For observers the average change of arousal of the therapist carried importance.

Similarly, research conducted by Vail (2023) also combines visual aspects such as body posture, facial representations of emotion, gaze patterns as well as verbal aspects to model therapist and client interaction.

These approaches are important to understanding therapeutic relationships, as they highlight that these interactions, as well as human interactions as a whole, are multifaceted and facilitated through a mixture of verbal and non-verbal aspects.

However, they also provide inspiration for attempting to better understand and possibly use linguistic features to predict working alliance, as both include some form of these in their modelling.

Additionally, Vollebregt (2023) notes that data which lends itself to such multimodal analyses is scarce, through poor video quality or obstructed faces. As such, methods which are based on transcriptions and audio may be more practically oriented, as this circumvents some of these issues.

2.7 Analysis

As laid out in the previous sections, the emotional content and linguistic alignment are the main features of interest for identifying the therapeutic working alliance. Table 1 summarises the main approaches taken for transcript analysis of dyadic therapy data.

	Emotion	Pos/Neg	Valence	Arousal	Alignment
Vollebregt (2023)	✓	✓	✓	✓	
Vail et al. (2018)	✓				
Vail et al. (2022)					✓
Negri et al. (2019)				✓	
Bayerl et al. (2022)					✓

Table 1. Review of previously used text analyses for WA prediction

Here, emotion refers to the specific emotions displayed in the transcribed speech of therapist and patient. As laid out by Vollebregt (2023), additional affective features such as the overall positivity and negativity, the sentiment, of conversations alongside arousal and valence can be good indicators to capture the emotional states of the participants of therapy sessions.

Alignment in Table 1 refers to the linguistic alignment previously discussed. Here it is important to note that the research of Vollebregt (2023) also delves into a type of synchrony, however there it refers to a visual alignment of facial action units (AUs), used in coordination with the facial action coding system (FACS) introduced by Ekman and Friesen (1978).

For specific emotions, the transformer model EmoRoBERTa, originating from the research of Kamath et al. (2022) was used by Vollebregt (2023) on English translations of Dutch therapy session transcripts. This model was used because of its capability of identifying 28 separate emotions, as well as a lack of a well-performing Dutch model for the same purpose. While an attempt was made to fine-tune the Dutch RobBERT model by Delobelle et al. (2020) to extract specific emotions directly from Dutch text, this was not successful due to a limited emotional Dutch dataset.

Vail et al. (2018) on the other hand used the Linguistic Inquiry and Word Count (LIWC) tool by Pennebaker et al. (2001), which is a dictionary matching tool for analysing emotions in text commonly used in psychological research, which will be detailed further in Chapter 3. While this tool has the advantage of not needing training or fine-tuning as well as being

translated into Dutch, its dictionary-based nature also represents limitations in dealing with word ambiguity and context.

For overall sentiment alongside valence and arousal, Vollebregt (2023) made use of the Valence Aware Dictionary and sEntiment Reasoner (VADER) by Hutto and Gilbert (2014) as well as XLM-RoBERTa, a model proposed by Conneau et al. (2019) and a sentiment oriented version of the RobBERT model mentioned previously. VADER is a dictionary-based tool with similarities to LIWC, while XLM-RoBERTa is another transformer model, the capabilities of which include valence and arousal prediction. In the work of Negri et al. (2019) an Italian-based dictionary was used to analyse the arousal, in a similar fashion to VADER and LIWC. It should be noted that VADER was chosen by Vollebregt (2023) for its high accuracy compared to other sentiment analysis tools such as TextBlob, developed by Loria et al. (2018). A downside to VADER is that its capabilities are focused on English.

The linguistic alignment analysis was handled in similar ways in the works of Vail et al. (2022) and Bayerl et al. (2022). Based on the original work of Müller-Frommeyer et al. (2019), a reciprocal linguistic style matching (rLSM) score is calculated at an utterance level. Importantly, Vail et al. (2022) demonstrates that LIWC can be used to calculate the rLSM score, increasing the relevancy of LIWC as a tool.

3. Methodology

3.1 Overview

The main tasks in this thesis are a transcription of the therapy video data, textual analysis focusing on affect and sentiment, followed by a prediction of working alliance using extracted features.

The transcription is be handled using WhisperX, see Bain et al. (2023), and PyAnnote, see Bredin et al. (2020). The inspiration for this comes from the research of Vollebregt (2023), who identified the Whisper model by Radford et al. (2023) as a high-performance tool for transcription. Since then WhisperX has become available, which will help with additional tasks such as diarization.

RobBERT, a Dutch language model by Delobelle et al. (2020), is fine-tuned to analyse arousal and valence. Furthermore, a pre-trained version of RobBERT is used to analyse the sentiment in text. The system used for fine-tuning as well as later correlation analyses utilises an Nvidia RTX 4090 GPU with 24GB of VRAM.

For correlation analysis and prediction the sklearn library is employed, see Pedregosa et al. (2011), from which linear regression, support vector regression and K-nearest neighbour models are used alongside leave-one-out cross-validation.

3.2 Data

The data used for this research originates from a study conducted by Bruijniks et al. (2020), in which the depression symptoms of patients were compared to the effects of attending therapy sessions, either cognitive behavioural therapy or interpersonal psychotherapy, once or twice weekly.

Cognitive behavioral therapy (CBT) is a form of psychotherapy that focuses on changing unhelpful cognitive distortions such as thoughts, beliefs and attitudes, as well as behaviors, as laid out by Beck (1979). This is done in an effort to improving emotional regulation and developing personal coping strategies that target solving current problems.

Interpersonal psychotherapy (IPT), on the other hand, emphasizes the interpersonal relationships of the patient. Described by Klerman et al. (1996), it's centered around the idea

that personal relationships can contribute to the onset or worsening of depression. Through this focus, IPT attempts to alleviate this condition by addressing problematic relationships as well as improving the communicative and problem-solving skills of the patient.

The data set consists of a total of 438 individual recorded sessions with 89 patients and 76 therapists, communicating in Dutch. The patients have a mean age of 37.85 years (+/- sd 12.26) with 61.5% being female. The ages of the therapists ranged between 25 to 61 with 81.6% being female. There are around 12-20 recordings per patient. Many of the recorded sessions include WAI-S and WAI-SRT scores provided by the patients and therapists, as well as expert observers. These WAI scores represent the ground truths for the later working alliance prediction. It is important to reiterate that the ratings of the patients, therapists and observers are not identical due to the subjective nature of the perception of working alliance. Therefore individual predictions of these scores will be important.

Here it is important to note that the dataset used for correlation analyses and prediction later on is multitudes smaller than the original number of recorded sessions. This is due to a number of steps each of which reduced the number of usable videos. The first of these is a matching process between the available WAI scores and videos. Since the predictions aim to investigate all three raters; therapists, observers and patients, an automatic matching process is devised to collect the video names for which the ratings of all three parties is available, resulting in only 52 videos. Of these 52, only 26 were eligible to be used after a diarization threshold explained later on, which are further reduced to only 12 due to missing data-point errors as well as the removal of duplicate candidate sessions. This removal refers to keeping the most recent session of each candidate in case that multiple sessions of one candidate are present, in order to ensure independent observations in the final dataset as well as avoiding possible over-fitting issues. The most recent session is chosen as it reflects the most up to date state of the therapeutic relationship.

The Medical Ethical Committee of VU Medical Centre Amsterdam (registration number 2014.337) agreed to the recording of this data set, alongside the full informed consent of the participating therapists and patients.

A dataset that is important to mention is one by Speed and Brysbaert (2023), which is publicly available and will be used to fine-tune a BERT model as mentioned in a later section. This dataset consists of approximately 20,000 Dutch words which are rated for their valence and arousal as well as six basic emotions; happiness, anger, fear, sadness, disgust and surprise based on the discrete emotional model by Ekman (1992). The ratings for arousal and valence were captured using a Likert scale ranging from 1 to 5 with 1

representing calm or negativity and 5 representing positivity and excitement. The specific emotions were also rated using a similar five-point Likert scale although in this case, it captures the strength of an emotion associated with a particular word (1 being "not at all" and 5 being "very much").

In addition to this Dutch word-level affect data, the EmotioNL dataset by De Bruyne et al. (2021), available upon request, will also be used for fine-tuning as will become apparent in a later section. This dataset contains around 2000 Dutch sentences in the form of evenly split Tweets and TV show captions, which are rated on the affect dimensions of arousal, valence and dominance as well as categorical emotion labels that fall within the discrete emotional model by Ekman (1992). The affect dimension ratings in this dataset are on a continuous scale between 0 and 1.

3.3 Automatic Transcription

Since the data is in video form, transcription is necessary to allow for further linguistic analysis to take place. The speech-to-text transcription is handled using the WhisperX library, a multilingual language transcription tool originating from Bain et al. (2023), making use of the Whisper transcription model. This tool is advantageous as it makes use of batch interference, allowing for the processing of multiple audio clips at the same time, as well as being able to transcribe Dutch, which is the spoken language in the available data.

Also, the Whisper model is a rather robust transcription model, having been trained in a supervised way on around 125,000 hours of English translation data as well as noisy speech data from 96 countries, adding another 680,000 hours of training data. This allows for Whisper to be rather accurate in transcription, with an average Word Error Rate (WER) of 8.81%, as opposed to a WER of 7.61-10.5% in human transcription, according to Radford et al. (2023).

$$WER = \frac{S + D + I}{N} \tag{3.1}$$

The equation for the WER, which is an established measure of transcription performance originating from Woodard and Nelson (1982), can be seen in Equation 3.1, where S represents the number of substituted words, D represents the number of deleted words, I represents the number of insertions. On the lower half of the fraction, N represents the number of words present.

The parameters for WhisperX are adapted from Vollebregt (2023), as optimal working parameters were found for a similarly limited computational environment and for the same therapy data. A key change however is that the newer "large-v3" pre-trained transcription model is used instead of "large-v1", with a batch size of 5. This model achieved a WER of 6.5% for a manually checked 10 minute segment of a therapy video.

As WhisperX itself consists of a number of pre-processing steps before actual transcription, elaboration on these steps seems appropriate. An outline of processes is displayed in Figure 1.

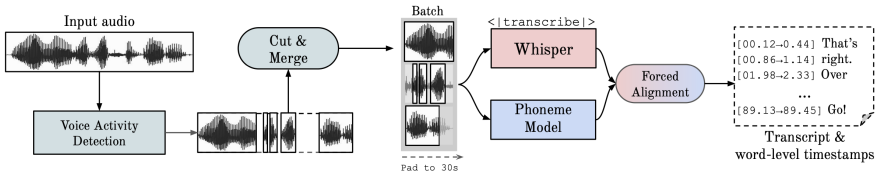


Figure 1. WhisperX process. Figure by Bain et al. (2023)

Starting with the raw input audio, voice activity detection (VAD) is applied to filter out background noise. This VAD is performed by a PyAnnote VAD model, see Bredin et al. (2020) and Bredin and Laurent (2021). On top of reducing noise, this step also already adds timestamps to the beginning and end of voice activity, allowing for reduced timestamp inaccuracy later on.

After applying VAD, the audio is then cut and merged. The segmented audio produced by the previous step is cut down into further segments based on their length, with any audio clip longer than thirty-seconds being split at points of low voice activity in order to reduce memory consumption in the Whisper model.

On the other hand, audio segments that are very short are merged together as they alone may not offer enough information or context for proper transcription to occur. Merging multiple adjacent short audio clips together, to a length close to the thirty-second maximum allows for these problems to be mostly overcome.

After both of these steps are complete, the remaining audio segments are then fed to the Whisper model in batches determined by the batch size parameter. The segments enter the model independently, meaning no context from previous segments is used, reducing possible hallucinations.

The Whisper model, proposed by Radford et al. (2023), is a transformer-based model. The main architecture is displayed in Figure 2. The previously mentioned thirty-second audio clips enter the Whisper model and are first re-sampled to 16kHz before being converted

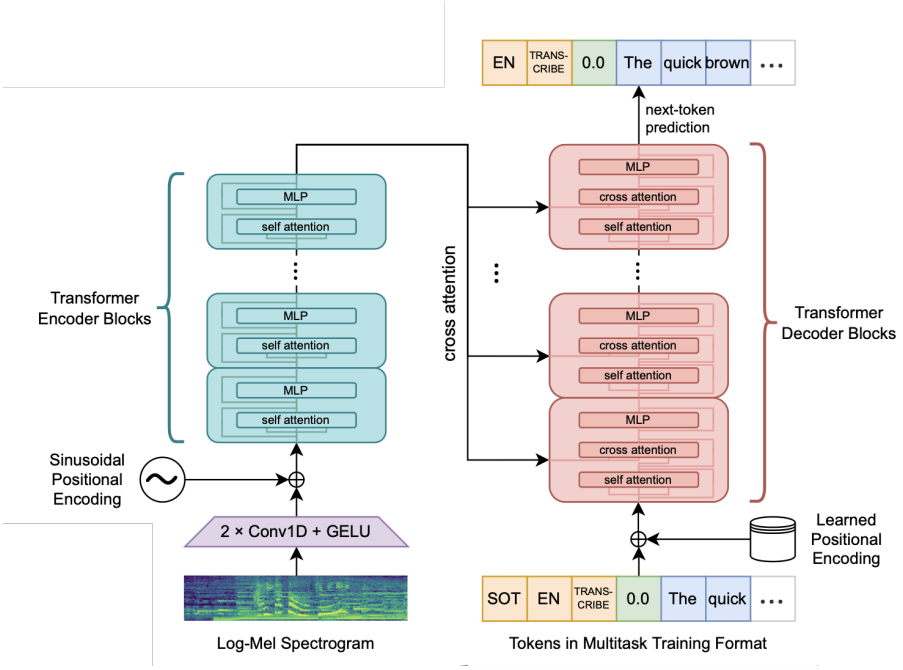


Figure 2. Whisper Architecture. Figure by Radford et al. (2023)

into a log-mel spectrogram format, a visual representation of the audio.

Due to the now visual nature of the audio, it is passed through two convolutional layers and a Gaussian Error Linear Unit (GELU) activation function, a non-linear activation function, which extract relevant features from the visual representation.

Positional encoding is then added to the output of the convolutional layers. Like the positional encoding mentioned in a previous section detailing the architecture of transformer models, here it is important as it allows for the positions of the spoken words to be remembered and used when generating output.

The encoder blocks in the Whisper model utilise multiple Long Short-Term Memory (LSTM) layers followed by a final layer normalisation before being passed to the decoder blocks which go on to predict the spoken words in text form, completing the transcription.

Finally, forced alignment is applied to the text segments produced by the Whisper model. This describes a process in which word-level timestamps are created using a phoneme model by identifying phonemes and matching these to the generated transcription text. The phoneme model in question is a Wav2Vec2 model, proposed by Baevski et al. (2020), which is also a transformer-based model.

Finally, the output of WhisperX is a word-level timestamped list of transcriptions. The nature of the data however dictates that these transcriptions stem from a conversation

between two people. For linguistic analysis, it is important that the transcripts of each are analysed individually. This is where speaker diarization is important, introduced in the next section.

3.4 Speaker diarization

According to Bredin et al. (2020), speaker diarization refers to the partitioning of audio or transcription as the one produced above into segments according to the speaker's identity. In the case of this research, this means that speaker diarization will segment the transcription into segments of therapist and patient utterances.

The process of speaker diarization is already part of WhisperX. The diarization model that is used is the "speaker-diarization-3.1" from the Pyannote library, see Bredin et al. (2020) and Bredin and Laurent (2021). For diarization, an established measure of performance is the diarization error rate (DER), used by the likes of Park et al. (2022a).

$$DER = \frac{F + M + C}{T} \tag{3.2}$$

Equation 3.2 shows the DER calculation, where F represents the amount of time for false alarms. Here false alarms refer to a false positive of voice activity detection, meaning that a segment was included that doesn't contain speech. M represents the amount of time for missed detection, where speech was not picked up. C represents the amount of time which is confused, meaning that the wrong wrong speaker is assigned to a segment of speech. T represents the total length of time analysed. For the same 10 minute segment mentioned earlier, a DER of 6.6% was achieved.

The process of speaker diarization shares an element with the earlier described process of transcription, namely the VAD system. Here it is used once again to filter out noise in the raw audio. PyAnnote then applies Speaker Change Detection (SCD) and Overlapped Speech Detection (OSD) simultaneously.

The outputs of these three processes are then used to transform the voice characteristics of individual speakers into vector form. Each segment of audio is then clustered and classified by finding the closest matching vector, leading to the final output of timestamps with speaker IDs.

Here it is important to note that the timestamps produced by WhisperX and the PyAnnote speaker diarization may not always be perfectly identical. Luckily WhisperX allows for a

re-alignment of timestamps, bypassing this problem.

While the automatic diarization assigns speaker IDs to the identified speakers of the videos, these IDs are arbitrary. This means that while for each video, IDs such as "SPEAKER_00" and "SPEAKER_01" are assigned, WhisperX has no way of identifying which of these IDs represents therapist or patient. For the purpose of speaker-level analysis however this distinction is quite important. In order to identify the speakers, a simple keyword matching function is developed. Here, the word "Sessie", which is Dutch for "Session", and the presence of question marks were deemed as signs of the therapist speaking. This keyword matching is very specific to the dataset at hand, as each recording starts with a therapist reading the session number. The presence of question marks being indicative of the therapist's identity was determined via manual inspection, as the therapy videos at hand include questions predominantly asked by therapists.

Unfortunately, there are performance issues associated with the automatic diarization, in which only one speaker is identified. These issues will be elaborated upon further in Chapter 4. To deal with these issues, a threshold is devised, based on speaker utterance ratios with a buffer to offer some lenience to therapy sessions that are less balanced in their conversation. Transcripts below this threshold are not used in down-stream tasks.

$$Threshold = 1.25 * |(\frac{Sp_1}{total}) - (\frac{Sp_2}{total})| \tag{3.3}$$

The final result of this transcription and diarization pipeline is therefore a speaker-level identified transcript including the start and end times of utterances, the spoken text, individually time-stamped words and a speaker identity of either "patient" or "therapist".

3.5 Transformers

Given that they are the basis of Whisper and BERT models, as introduced by Devlin et al. (2018), it seems appropriate to provide context about transformers.

First introduced by Vaswani et al. (2017), transformers are a relatively new model architecture, mainly used for natural language processing (NLP), which rely on an attention mechanism to infer input and output dependencies across input sequences rather than being limited to a context window like previous solutions including recurrent neural networks and long short-term memory models.

Figure 3 gives an overview of a transformer's architecture, divided into two visually

separate units, namely the encoder seen on the left-hand side and the decoder on the right.

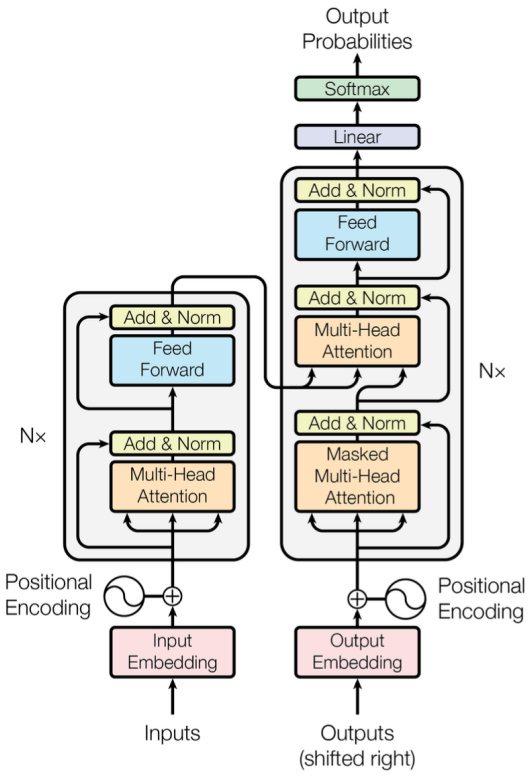


Figure 3. Transformer architecture - Figure by Vaswani et al. (2017)

The encoder serves the role of converting input sequences to continuous, context-aware representations which are later used by the decoder to generate an output sequence. This process starts with the input embedding layer, in which each word in an input sequence is mapped to a representational vector of numbers. Since the transformer does not incorporate recurrence, positional encoding is used to add positional information to each word vector. This is done by generating positional vectors using Equations 3.4 and 3.5 before adding them to the word representation vectors. Here, pos represents the position of a word in the input sequence while i represents a dimension in which the current sinusoid is located. The d_{model} represents the dimensionality of the input embeddings. Interestingly, there are two equations, which can be explained by the fact that Equation 3.4 is used for words at an even position, while Equation 3.5 is used for words at an odd position. This design choice allows for the positional encoding vectors to represent a wide range of positional relationships with minimal overlap, allowing for more rich embeddings.

$$PE_{(pos,2i)} = sin(\frac{pos}{10000^{2i/d_{model}}}) \tag{3.4}$$

$$PE_{(pos,2i+1)} = cos(\frac{pos}{10000^{2i/d_{model}}}) \tag{3.5}$$

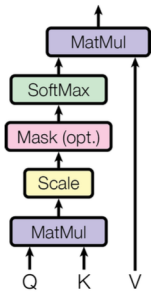


Figure 4. Scaled dot-product attention - Figure by Vaswani et al. (2017)

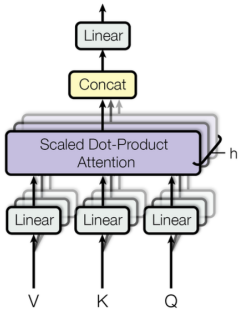


Figure 5. Multi-head attention - Figure by Vaswani et al. (2017)

Located in the actual encoder layer are two main modules. The first of these is a multi-headed attention module, which can be seen in greater detail in Figure 5. Here it is important to note the scaled dot-product attention component, which is the core mechanism of a transformer's self-attention and can be seen in Figure 4.

Starting with the scaled dot-product attention, the input embeddings created in the steps described above are first linearly transformed into three different sets of vectors; queries (Q), keys (K) and values (V). Here, the query vectors represent the positions in the input sequence to which other positions will attend, the key vectors capture the level of attention that each position should pay to other positions and the value vectors contain information about the original embeddings which is accumulated depending on the attention scores provided by the queries and keys, as seen in Equation 3.6, which mirrors the structure displayed in Figure 4. Here, the queries and keys undergo a dot-product matrix multiplication, denoted by the QK^T in which K^T denotes the transposed versions of K, before being scaled via a division through the square root of the dimensionality of the queries and keys which is represented by d_k.

The created attention matrix is then adjusted using the softmax function, meaning that high attention weights are endorsed while low scores are inhibited. These attention weights are then combined with the values through matrix multiplication which results in the output of the scaled dot-product attention component.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3.6}$$

The multi-head attention component works by first splitting the query, key and value

vectors into N vectors before going through the scaled dot-product process. This splitting creates multiple attention heads, as seen in Figure 5 as h, which are then concatenated and passed through a final linear layer.

The reason for this multi-headed approach is that each attention head should learn something different, resulting in richer final representations. Equation 3.7 demonstrates how the concatenation is handled, in which W_i^Q, W_i^K, W_i^V represent learnable weight matrices of the individual queries, keys and values of each head, while W^O represents a final weight matrix used in the concatenation.

$$\begin{aligned} MultiHead(Q, K, V) &= Concat(head_1, ..., head_h)W^O \\ \text{where } head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{3.7}$$

Finally, a feed-forward layer is included, described by Equation 3.8. This layer introduces non-linearity to the model through a ReLU activation function, represented by the $max(0, xW_1 + b_1)$ part of the Equation 3.8, allowing for more complex relationships to be captured. The input vector is x, while W_n and b_n represent the weights and biases of two linear transformations that occur.

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \tag{3.8}$$

It is important to note that each component, be it a feed-forward layer or a multi-headed attention layer, has a residual connection that facilitates a flow of information which propagates and preserves important information, as well as helping minimise the vanishing gradient problem. These residual connections are always normalised to stabilise learning by not allowing the connections to have too large of an impact.

The structure of the decoder of the transformer, seen on the right-hand side of Figure 3 is comprised of the same components as the encoder, except for an additional multi-headed attention layer, which takes keys and values from the encoder output, allowing the decoder to utilise information from across the input sequence when generating outputs.

The final linear transformation and softmax layer then allow the transformer to provide probability distributions for possible words in the decoder's output sequence.

3.6 Affect Analysis

In order to analyse the affect displayed in the Dutch transcripts, a custom language model is developed and implemented. This approach is based on RobBERT, a state-of-the-art Dutch language model by Delobelle et al. (2020), which itself is based on the BERT architecture. This model is chosen because it is a Dutch language model which has previously been successfully used in a number of natural language processing tasks such as sentiment analysis and coreference resolution, both in the original paper of Delobelle et al. (2020). The inspiration to fine-tune this model instead of using an off-the-shelf counterpart comes from the lack of a Dutch only affect analysis model, as well as an interest in a specific output of arousal and valence values from one model in order to aid the research of Dr. Sanne Bruijniks. The creation of a custom version of RobBERT is achieved in two main steps: Word-level pre-training and sentence-level fine-tuning.

Two datasets are used to train the custom RobBERT model. The first is a merged version of the word-level dataset of Speed and Brysbaert (2023), containing around 20,000 Dutch words rated for their arousal and valence scores. Originally, this dataset is comprised of separate files for arousal and valence ratings, which are merged for training. Additionally, the affect scores are normalised to fall within a 0 to 1 range to match the format of the sentence-level dataset and the desired output of the model.

The second is a merged version sentence-level dataset of De Bruyne et al. (2021) containing around 2000 Dutch sentences also rated for their arousal and valence. This dataset originally contains separate files for affect rated tweets and captions. Due to the relatively small size of this dataset, these files are merged and then shuffled before being split into a training set consisting of 80% for use with cross-validation and a final validation set of 20%.

For the first step in the custom model creation, the RobBERT model "robbert-2023-Dutch-base" is pre-trained on the word-level dataset described above. This step aims to provide the model with a foundational understanding of emotional valence and arousal in Dutch vocabulary. Such word-level pre-training has been previously linked to higher performances in sentence-level tasks of similar BERT based language models Howard and Ruder (2018). This process involved splitting the word-level dataset into a training set consisting of 90% and an evaluation set of 10% and training the model to predict both arousal and valence simultaneously.

Following this word-level pre-training, the model was then fine-tuned on a combined and shuffled version of the EmotioNL dataset by De Bruyne et al. (2021). This step aims to

help the model integrate word-level understanding into more applicable sentence-level affect predictions. The fine-tuning process employs a 5-fold cross-validation, striving for a robust performance while mitigating over-fitting. This approach also seeks to maximise the utility of the relatively small sentence-level training dataset.

The main model architecture is based on the RobertaForSequenceClassification class from the Transformers library. This class was modified to output two continuous arousal and valence values instead of discrete classes, treating the task as a multi-output regression problem. In addition to this, several task-specific modifications were made including a custom PyTorch Dataset class, which was implemented to handle the desired multi-output, encapsulating the logic for tokenizing input texts and pairing them with their corresponding arousal and valence labels, as well as a custom Trainer class extending the HuggingFace Trainer by adding functionality to track and store training and evaluation losses.

The final model was evaluated on a held-out validation set containing 20% of the original shuffled EmotioNL dataset, providing an unbiased evaluation of the model's performance on unseen data. In addition to this, several Dutch sentences that were reviewed by a Dutch speaker were used as manual verification of the model's ability to predict arousal and valence, which can be seen in Chapter 4.

In addition to this custom RobBERT model, the sentiment present in text is also analysed. This task however is achieved using a pre-trained version of the RobBERT Dutch model called "robbert-v2-Dutch-sentiment" originating also from the work of Delobelle et al. (2020). The sentiment is captured using the binary labels of either "positive" or "negative".

3.7 Construction of feature set

Here it seems useful to reiterate the pipeline of analysis. First, the therapy videos are converted to audio files which are automatically transcribed and diarized resulting in a time-stamped transcript of each session. The speakers are then identified as either "patient" or "therapist", leading to identified versions of the transcripts. The text of transcripts is then analysed on a speaker-level using the fine-tuned RobBERT model to extract the arousal and valence. The pre-trained RobBERT sentiment model also used to extract the sentiment, also at a speaker-level. This process results in a data frame per transcript containing the start and end times of a speaker's utterance, the speaker's identity in the previously described "patient" or "therapist" format, a sentiment label of either "positive" or "negative", and finally separate numerical values on a scale of 0 to 1 for arousal and valence.

While the speaker-level granularity of the affect and sentiment analysis allow for useful visualisations of the changes of these features throughout a session, it does create a high-dimensional feature space per speaker per transcript. Therefore, in an effort to reduce the raw information to a set of meaningful, descriptive features, the following values are calculated.

Firstly, the averages of arousal and valence. These provide an overall representation of the emotional states of the participants during the session. In addition to the averages, the minimum and maximum values of arousal and valence per speaker are also extracted, as they point towards emotional extremes in the sessions.

Secondly, the variances of arousal and valence, which capture the fluctuation of these affect dimensions. This is a seemingly important feature, as it better describes the dynamic emotional states of therapy participants as opposed to a simple average. Additionally a high variance in affect can indicate emotional flexibility which has been associated with psychological health by Kashdan and Rottenberg (2010), who note that disorders such as depression feature a loss in emotional flexibility.

Temporal features are also collected. These include the total speaking times for both participants, their total number of utterances, as well as the average response times of patient and therapist. While manual inspection of the therapy sessions suggest that both the therapist and patient are somewhat equal in their conversational participation due to the question and answer structure often associated with therapeutic sessions, it would be interesting to see if imbalances correlate with working alliance scores. The average response times may also carry some significance as quicker times have previously been linked to a higher sense of social connection by Templeton et al. (2022).

Since the analysed transcripts include temporal data such as the start and end times of speaker utterances, the affect and sentiment analysis results can be interpreted as time-series data. This representation allows for synchrony measures to be calculated between the arousal, valence and sentiment of therapist and patient. This is achieved by creating a common time grid with one-second intervals spanning the entirety of a session. The affect and sentiment data is then resampled to this grid using a forward-fill method in which the last known value is extended until a new value is available. For this resampling, mid points in time for utterances are found as half the summation of their start and end times. An example can be seen in Appendix A.3. The synchrony of each measure is then calculated using Pearson's correlation coefficient between the corresponding therapist and patient time-series. The resulting synchrony values range between -1 and 1, indicating the degree and direction of co-variation between the affect and sentiment throughout a session.

3.8 Prediction

To investigate the relationships between the extracted features and the working alliance scores, as well as their sub-components, correlations and predictive models are investigated. Here it is important to note that the target variables are composed of an overall working alliance score as well as individual scores for each sub-component of this working alliance score; task, goal and bond. The overall working alliance score falls into a range of 0 to 60, while the individual scores fall into a range of 0 to 20 each.

Initially, Pearson correlation coefficients are calculated. This is done between each extracted feature and the WAI scores of therapist, patient and observer in an effort to identify potential relationships. These correlations are calculated using the scipy.stats module before being sorted by the absolute value of the coefficients to highlight the strongest among them.

Following the correlation analysis, three models are trained to predict the various working alliance scores and sub-components. These models are a linear regression model, a support vector regression model and a K-nearest neighbours model, chosen for their ability to handle small datasets.

The models are trained and evaluated using leave-one-out cross-validation. In this cross-validation, the models are trained repeatedly on nearly the full final dataset, each time leaving out one sample that acts as the test set. This process is repeated until each sample has been left out once. Then, the recorded performance scores, in form of the root mean-squared error, are averaged to capture the overall performances of the models. It is important to note that R-squared values can not be calculated due to only one sample being in the test set.

Additionally, an extra model is created to act as a baseline for comparing the model performances. This extra model simply predicts the mean of the target variable. This creates a benchmark to judge whether the other models actually manage to predict the target values.

4. Results

4.1 Diarization issues

Due to the sensitive nature of the therapy videos, excerpts from transcripts as examples of detected issues can not be included. This section therefore relies on more descriptive results in an effort to illustrate the issues surrounding the diarization process. It is also important to note that this section is representative of experimental results, as most findings are based on manual inspection.

While WhisperX is currently a seemingly good tool for automatic transcription and diarization, it is not certainly not flawless. Despite the established WER and DER performance measures, which suggested that the performance of the WhisperX pipeline on real data, meaning use outside of near perfect test audio conditions, is relatively good, it quickly became apparent that the performance can vary much more drastically than previously noted.

This lack of performance presented itself through the manual inspection of some of the transcripts. Since the step after the automatic transcription and diarization is to identify and effectively split the transcripts into the parts of therapist and patient, it seemed unusual that for some transcripts this was either not possible or resulted in very few sentences for either of the two speakers.

Such transcripts were identified as problematic through further manual inspection. Viewing the videos attributed to these transcripts revealed that it was indeed the automatic diarization failing, as they consisted of the usual setup of a patient actively talking to a therapist, ruling out the possibility that either party was simply not speaking.

The manual inspection did however lead to the hypothesis that audio quality may be a limiting factor for the automatic diarization, as this seemed much lower in the videos associated with the problematic transcripts. Harsher background noises such as a running fan were also identified. In an effort to overcome these issues, attempts were made to enhance the audio.

The first attempt was made using the noisereduce library by Sainburg (2019), which leverages spectral gating, a threshold based on the spectrogram of a signal such as speech, to remove noise from audio. Unfortunately, while further manual inspection of the cleaned

Session	Overall quality	Speech quality	Background noise quality
1015 session 4	2.140	2.783	2.708
7013 session 8	1.260	1.585	1.361
9008 session 4	1.401	1.856	1.739

Table 2. SpeechMOS results

audio suggested that it was less noisy, there were no improvements in the diarization.

As noise reduction did not seem to have an impact on the diarization, an attempt at speech enhancement was made. For this, a state-of-the-art pre-trained speech enhancement model by Ravanelli et al. (2021) was used, namely the "sepformer-dns4-16k-enhancement" model. This model utilises a transformer based architecture to process speech, generate masks to isolate speech and return an enhanced, clarified version of the input audio. Again manual inspection suggested that the enhancement did work as intended, but the diarization did not improve and even the transcription suffered, causing multiple instances of the same wrong sentences being repeated in the inspected transcripts.

As these enhancements did not seem to alleviate the issue, a final investigation was made into the audio quality via the use of the "dnsmos" model of the SpeechMOS library, associated with research by Reddy et al. (2022). This convolutional neural network model is a speech quality assessment tool designed to predict human ratings of speech quality. Table 2 shows the measures of overall, speech and background noise quality. Each of these scores exists on a scale of 1 to 5 with 5 being the best.

In Table 2 three therapy sessions are presented. The first, 1015 session 4, represents a session associated with a well formed transcript that was largely diarized correctly, within the WhisperX diarization error rate. The second, 7013 session 8, is also associated with a good transcript. The last however, 9008 session 4, is not, with the diarization resulting in only one speaker identity being identified.

Interestingly, the results of the audio quality assessment in Table 2 shows that the scores of session 7013 are lower in all three areas when compared to 9008, suggesting that audio quality may not be the cause of the bad diarization after all.

This realisation led to further, different approaches to improve diarization. One such approach was to replace the Pyannote diarization model used in WhisperX with fine-tuned versions. The first of these, uploaded to Huggingface by Kamilakesbi (2024b) is fine-

tuned on the English version of the CallHome dataset from Canavan et al. (1997a). The second was also uploaded by Kamilakesbi (2024a) and is fine-tuned on the German version of the CallHome dataset by Canavan et al. (1997b). The CallHome datasets consist of audio-transcript pairs of telephone conversations. The English and German versions were chosen based on their linguistic similarity to Dutch, while the telephone conversations represented more challenging conditions for diarization. Unfortunately neither of these models improved the diarization for the therapy sessions at hand.

A last attempt was therefore made to replace the Pyannote diarization model with one from the Nvidia NeMo framework, see Park et al. (2022b), however time-limitations and technical issues did not lead to success in this endeavour.

4.2 Affect analysis model

As described in Chapter 3, a RobBERT model was fine-tuned to predict arousal and valence from Dutch text. Several metrics were calculated to evaluate the performance of this custom RobBERT model. These include the mean squared error (MSE) which measures the average squared difference between predicted and actual values, the root mean squared error (RMSE) which presents the MSE measures in a format matching the output variables, the mean absolute error (MAE) which describes the absolute difference between predicted and actual values, the R-squared value representing the variance in the dependent variable explainable by the independent variable, a measure for accuracy which in this case represents the proportion of predictions which land within 0.1 of the true values and finally Pearson's correlation coefficient (r) which is a measure of the linear correlation between the predicted values and the ground truth values.

Metric	Without pre-training	With pre-training
MSE	0.030	0.029
RMSE	0.173	0.169
MAE	0.140	0.132
R^2	0.336	0.360
Accuracy (within 0.1)	0.421	0.458
Pearson's r Arousal	0.627	0.614
Pearson's r Valence	0.667	0.701

Table 3. Evaluation metrics of fine-tuned RobBERT models

Here it is important to note that the RobBERT model with pre-training refers to the fine-tuned RobBERT model as described in Chapter 3, while the model without pre-training refers to the same fine-tuning approach without the addition of the word-level pre-training step. Evidently, both approaches perform similarly, with the word-level pretrained model performing marginally better overall, with a slight decrease in its predictive power for arousal. The model with pre-training is nonetheless chosen as the main model for the rest of the analysis and therefore will be the model the rest of this section refers to.

Looking at the results in Table 3, the metrics of primary importance are the RMSE of 0.169 and the MAE of 0.132 of the model with pre-training. Both of these metrics are measured on a scale of 0 to 1. Their values indicate that on average the fine-tuned model's predictions deviate by around $0.13 - 0.17$ from true values. While such a level of error is not negligible, it suggests that the model has captured meaningful patterns in the training data.

Additionally, it is important to note that in the context of emotional labeling and prediction, human raters often disagree about affect ratings. This is demonstrated in the text of Speed and Brysbaert (2023), in which ratings per word were collected from at least 8 individuals, the number of whom was increased in case the reliability of the raters was below 0.8. This reliability refers to an interclass correlation coefficient used to assess the consistency of ratings among multiple raters.

It should be noted that this accuracy metric is designed to help understand the model's performance, as accuracy as a measure is predominantly associated with classification rather than regression problems. Here it is used to capture the percentage of predictions that fall within a seemingly acceptable threshold. The accuracy metric shows that 45.75% of predictions fall within 0.1 of the true value. This accuracy value holds promise, as the model manages to achieve this accuracy while the RMSE and MAE both lie above 0.1. This further supports the model's ability to approximate arousal and valence levels, especially considering the continuous nature of these ratings.

The R-squared value of 0.360 indicates that the fine-tuned model explains approximately 36% of the variance in target variables. While low in comparison to some predictive tasks, this still seems reasonable considering the complexity of human emotion. Additionally, the relationship between text and affect is certainly not linear, and a low R-squared value can still provide valuable insights into trends of affect.

For the Pearson correlation coefficients, the value for arousal of 0.614 suggests that there is a moderate positive correlation between the predicted values of the model and the

Sentence	Arousal	Predicted Arousal	Valence	Predicted Valence
Het concert van gisteravond was waanzinnig energiek en spectaculair!	0.900	0.869	0.900	0.836
Het regent zachtjes buiten terwijl ik een boek lees.	0.200	0.310	0.700	0.577
Ik ben ontzettend blij met mijn nieuwe baan!	0.700	0.710	0.900	0.848
Het verlies van mijn huisdier heeft me diep bedroefd.	0.300	0.382	0.100	0.145

Table 4. Manual evaluation sentences for fine-tuned RobBERT model

ground truth labels. For valence the value of 0.701 is higher, suggesting a strong positive correlation. This means that the model's outputs align reasonably well with the ground truths, especially as the overall error metrics of MSE, RMSE and MAE indicate relatively low prediction errors.

In addition to testing the fine-tuned model on an unseen part of the EmotioNL dataset, four additional sentences were used to verify its ability to judge affect. These four sentences, seen in Table 4 were first generated alongside arousal and valence scores using ChatGPT4-o by OpenAI (2024) and additionally vetted by a Dutch speaker. While this step should not be seen as an entirely robust verification, it does help to illustrate the model's capability to capture affect in Dutch text.

The fine-tuned RobBERT model is used to extract affect information from the therapy session transcripts, as explained in Chapter 3. Figure 6 provides a temporal visualisation of the affect of a patient. The therapist's perspective of this example can be seen in Figure 9 in Appendix A.3

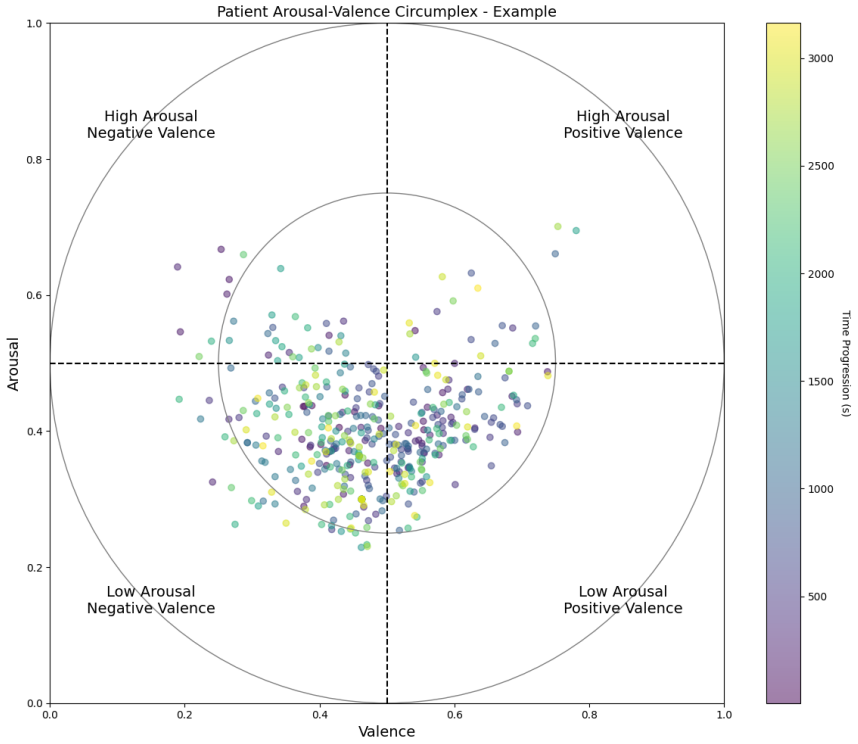


Figure 6. Example of a temporal affect circumplex model of a patient

4.3 Correlation analyses and predictions

As mentioned in Chapter 3, a number of features were calculated for correlation analyses and predictions. For each speaker, the averages, variances, minimum and maximum values of arousal and valence, as well as the total speaking times, number of utterances and average response times were extracted. In addition to these, the synchrony between speakers in arousal, valence and sentiment were also extracted.

First, a descriptive look into the final dataset is useful, given its small sample size. The distributions of WAI scores for patients, therapists and observers can be seen in Figures 11, 12 and 13 in Appendix A.4. The distributions of patient related features can be seen in Figure 14. Figure 15 shows the distributions of therapist related features, with the remaining features being displayed in Figure 16.

The patient rated WAI ratings demonstrate the widest range and variability among the three raters. In each of the components as well as the WAI score itself, there seem to be bimodal distributions, which could suggest the existence of subgroups within the patient experiences. The therapist rated WAI ratings on the other hand cluster more tightly while showing peaks in the upper ends of the bond and goal component ratings. The observer ratings generally seem to display near-normal distributions, especially for the overall WAI

Combination	Correlation	p_value
Observer task & Patient avg valence	0.799	0.002
Therapist goal & Patient avg valence	0.633	0.027
Patient bond & Patient pos sentiment ratio	0.633	0.027
Patient bond & Patient neg sentiment ratio	-0.633	0.027
Observer WAI & Patient avg valence	0.627	0.029
Therapist goal & Therapist min arousal	0.605	0.037
Therapist task & Therapist arousal variance	-0.602	0.038
Therapist bond & Therapist num utterances	0.590	0.043
Observer goal & Therapist valence variance	-0.587	0.045

Table 5. Correlations with significant p_values

score, except for the task component, which exhibits a peak almost moving towards a bimodal distrbution.

Examining the distributions of the features reveals further interesting insights. The arousal and valence measures seem to display generally skewed distributions, especially for the variance measures for both patients and therapists. Here it seems as though patients tend to display higher variances of arousal than therapists, while the reverse seems to be the case for valence. For time related features such as the total times spoken by therapist and patient, it seems as though patients generally dominated the sessions.

Moving forward, Table 5 highlights the most important results of the correlation analysis. The displayed feature-target combinations represent the correlations with statistical significance, denoted by a p_value below 0.05.

Looking at Table 5, a seemingly important feature is the average valence of the patient. It displays strong positive correlations with the goal component of a therapist's working alliance rating, as well as the task component and overall score of the observer, suggesting that patients who communicate in a emotionally positive way tend to be associated with higher alliance ratings in these avenues.

An interesting aspect is captured by the symmetrical correlation found with a patient's sentiment. A higher positive sentiment ratio is correlated with a higher patient rating of the bond component, while the reverse is true for a negative sentiment ratio, suggesting that patients who express more positive sentiment in a session rate their bond with their

therapist higher. Unfortunately, this symmetry also presents some problems, which will be discussed in Chapter 5.

The minimum arousal of the therapist seems to correlate with the goal component from the therapist's perspective. A higher minimum arousal suggests a higher rating of the goal component, while a lower variability of arousal seems to suggest a higher rating of the therapist's task component, suggesting that their own emotional state plays a role in their perception of alliance.

Additionally, a therapist's number of utterances is correlated positively with their perception of the bond component, which could be a reflection of their engagement with the patient. Finally, a therapist's variance of valence is negatively correlated with the observer's goal component, which may suggest that emotional stability reflects better on goal agreement to an outsider.

Despite the small sample size of the final dataset, prediction of the working alliance scores is attempted using a number of regression models. In Table 6, Mean refers to the model always predicting the mean value of the target which serves as a performance baseline, Lin Reg refers to a linear regression model, SVR refers to a support vector regression model and KNN refers to a K-nearest neighbour model. The Table itself contains the

Target	Mean	Lin Reg	SVR	KNN	Best K
Patient bond	2.583	7.439	3.553	2.833	3
Patient goal	2.333	7.658	2.236	2.416	4
Patient task	2.417	7.492	2.480	2.333	1
Patient WAI	6.444	21.034	6.528	5.833	1
Therapist bond	1.583	8.948	1.517	1.617	5
Therapist goal	1.153	6.469	1.303	1.292	2
Therapist task	1.056	5.457	1.249	1.125	2
Therapist WAI	3.194	20.874	3.630	4.083	5
Observer bond	2.417	7.702	2.854	2.306	3
Observer goal	1.333	1.614	1.688	1.639	3
Observer task	1.431	3.078	1.936	1.458	4
Observer WAI	4.417	10.712	5.584	4.458	4

Table 6. RMSE performance measures for different models

RMSE performance measures of each model, as well as the best K value for the K-nearest neighbour regression model. A full view of the K-nearest neighbour performances can be seen in Appendix A.5.

Looking at Table 6, the K-nearest neighbours model generally outperforms both the linear regression model and the support vector regression model. Overall it demonstrates consistently low RMSE values, although its performance generally still lacks behind the simple mean predicting model, except for the patient's rating of the overall WAI score and the task component, as well as the bond component of the observer. Interestingly, the linear regression model is consistently the worst performing model, which may suggest that non-linear relationships exist.

In terms of the different raters; patient, therapist and observer, the RMSE values for patient score predictions are generally higher than for the other raters. The RMSE values when predicting the therapist ratings are generally the lowest, while the performance for predicting the observer ratings are comparable but slightly higher.

5. Discussion

5.1 Diarization

As laid out in Chapter 4, the automatic diarization process is not yet guaranteed to perform at a consistently high level. Although efforts were made to alleviate these issues, ultimately these were not yet fruitful, leading to the use of the custom threshold seen in Equation 3.3 to filter transcripts.

This area presents a large source of future research direction, as reliable improvements to transcription and diarizations would aid the speaker-level analysis greatly. Being able to accurately divide conversational partners in transcripts allows for deeper insights into the parts they play.

Improvements here would also serve the interest of automatic transcription as a whole, as a key motivation is to reduce the resources needed for transcription. Ideally human transcription would solve a lot of the issues presented in Chapter 4, however this approach is resource intensive and not always a possibility, whether it be due to research team size or data sensitivity.

A possible approach to improving the diarization specifically for the dataset explored in this research is to fine-tune the Pyannote diarization model independently. Such possibility is presented by Akesbi and Gandhi (2024) in the form of a repository for fine-tuning speaker diarization models. Here it is suggested that around 10 hours of audio with human-transcribed text can be used to fine-tune diarization models.

5.2 Affect analysis

The fine-tuned RobBERT model demonstrates considerable utility for detecting the general trends of affect in Dutch text. While it may not perfectly predict individual affect scores, its ability to capture overall patterns and changes in affect when used in speaker-level analyses of transcripts can provide valuable insights.

While human rating is still likely to be a more robust approach when seeking accurate affect ratings, the strength of the fine-tuned model lies in its efficiency and consistency.

Despite the presented performance, the fine-tuned model carries with it several limitations

that need to be acknowledged. The first is the unexplained variance demonstrated by the R-squared value. At 0.336, it suggests that a significant portion of variance in affect is not captured by the model. This could be due to a number of factors such as the inherent difficulty in quantifying emotions, limitations in the model's architecture or training process or inconsistencies in the training data. Since the model was fine-tuned on the relatively small EmotioNL dataset, it is likely that a larger dataset of Dutch sentences with affect ratings would allow for a higher overall performance of the model.

Future work surrounding affect prediction could focus on several areas. The first and one of obvious importance is data expansion. While the existence of EmotioNL and the individual word ratings dataset provide a seemingly decent foundation to capturing affect in the Dutch language, larger datasets of rated Dutch sentences could improve the model performance.

Another possible area of future work could focus on long-range dependencies. Currently the model is used to predict affect scores of text segments independently. This means that the context of a conversation is not fully utilised. Better handling of such conversation-wide context could improve the model's capability of understanding and predicting affect in extended dialogues.

5.3 Correlations and predictions

As seen in Chapter 4, the predictive powers of linear regression, support vector regression and K-nearest neighbour models were low, not consistently beating the baseline model predicting the target variable means. While this lack of performance may suggest that the prediction of working alliance using the methods presented is not a viable strategy, there may also be other explanations.

A likely significant factor in the low predictive power is the simple fact that the sample size of the final feature dataset is very small. With only 12 samples, feature-sets collected from therapy sessions that passed through the proposed pipeline, it is very likely that there is simply insufficient data to reliably model the relationships between variables.

This low amount also points to the likelihood of over-fitting, which may explain why the linear regression model performs much worse than the other models in Table 6. Furthermore, any measurement errors or outliers are magnified with such low sample size. This applies to both the collected features as well as the ground truth working alliance rating. It is also important to highlight the fact that the final model selection and reported performance results in Table 6 are based solely on leave-one-out cross-validation. An

evaluation set was not used due to the small sample size of the final dataset, meaning the performance measures are likely to be overfit and have limited generalizability.

Here it is useful to refer back to the WAI rating distributions. As described earlier, the overall performances of all models were higher when predicting the scores of therapists and observers. This likely reflects the bimodal distributions of the patient WAI scores displayed in Figure 11, as unimodal distributions are easier to model. Additionally, this would explain why the mean predicting model also performs worse, as by predicting the mean, it may be predicting values between the distribution peaks, leading to a higher error.

This low sample size also leads to problems with the dimensionality of features. While the extracted features were designed to reduce the number of data points per transcript, their number is still higher than the number of available samples, increasing the risk of spurious correlations. It is therefore also unlikely that the present predictive models would generalise well, even if their predictive power was higher, as the small sample size is unlikely to represent the full spectrum of therapy dynamics. This combination of low sample size and high number of features is also likely the reason why the baseline model predicting the mean scores outperforms the others, as there is not enough data to properly model any trends or relationships properly.

Another issue that was highlighted in Chapter 4 is the fact that some features such as the positive and negative sentiment ratios form symmetrical correlations to target variables. While the correlations shown in Table 5 reinforce the idea that sentiment has an impact on a patient's perception of the therapeutic bond, such symmetries also indicate redundant information, as the positive and negative ratios seem to provide the same information from opposite perspectives. Again the low sample size gets in the way of deeper analysis into the effects and importance of each feature, but future research could focus on feature selection and reduction, given more data is available.

In addition to these drawbacks, it is also important to take the complex nature of therapeutic working alliance into account. While emotions likely play a part in the establishment and experience of such an alliance, it is also likely that the presently collected features do not fully represent the all factors that play a role in its upkeep.

5.4 Multimodal machine learning

While the presented analysis pipeline and predictions are more akin to a unimodal approach, the insights gained here still offer value to future multimodal research into therapy sessions and vice versa.

Firstly, the speaker-level granularity may allow for deeper insights into the the individual contributions of the therapy participants when combined with analyses based in modalities outside of text. An example for this is using the speaker identified transcript timestamps as a base for audio analysis. It should be noted that such speaker-level audio analysis was conducted based on the present audio analysis of Vollebregt (2023) for Dr. Sanne Bruijniks in the course of this project. Due to time constraints and the already present small final dataset with feature dimensionality issues, it was decided not to include this vein of analysis here.

Secondly, the issues associated with diarization may be alleviated in part through the use of visual features. For example, in a controlled setting, with good camera angles, speakers could be identified based on their facial movements, allowing for a more solid prediction of who is talking at a time.

5.5 Research questions

Research question 1: What linguistic features are relevant to the therapeutic working alliance?

From previous research, it seems that the use of emotional language, here captured through the affect dimensions of arousal and valence, are of main importance. This is reflected in part by the found feature correlations. This however should not yet be accepted as solid fact and the limitations of this study kept in mind. It is important to highlight that the features collected here are unlikely to be fully representative of all relevant linguistic features.

This is especially true when considering the nature of the three sub-components of the working alliance inventory; bond, goals and tasks. While this research focused mainly on the dimensions of affect, there are likely more elements at play outside of emotion and sentiment. For example, the goal and task components might be more predictable using specific language elements. Future work could therefore delve into identifying agreement between therapists and patients on tasks and goals. An early attempt at this was also made using the Linguistic Inquiry and Word Count (LIWC) tool developed by Pennebaker et al. (2001). Due to time limited time, this attempt did not yet lead to useful results and was therefore left out of the presented pipeline.

Research question 2: How can linguistic features be extracted from video data?

This question ended up being central to the analysis pipeline presented in this research.

While there are certainly flaws in the automatic transcription and diarization, as well as room for improvement in the affect extraction model, it does represent a proof of concept. Extraction of linguistic features from video and audio is clearly possible. Future work should focus on improving aspects such as reliable diarization for real life audio environments which should increase the effectiveness of these methods.

Research question 3: To what extent can linguistic features be used to predict this working alliance?

This question is not answered in full. This is due to the fact that the final dataset of features was simply too small to make solid predictions. Future research should therefore seek to improve the predictive models by expanding the final dataset. This should be possible through the creation of high quality input data and improvements to the automatic diarization.

Despite the lack of solid predictions, significant correlations are found between working alliance scores, their components and several extracted features. Future research should therefore aim to investigate these correlations further, as they may lead to more solid predictions with more data available.

It is also important to consider that a simpler prediction target may improve the performance of the predictive models. Currently, these models aim to predict a score on the same scale as the ground truths. However, simplifying this prediction to a more binary prediction such as high and low working alliance could improve the performance and usefulness of these models. In any case, more data is needed.

6. Conclusion

To conclude, the methods described here can be used as a proof of concept. Automatic transcription, while needing improvement, can lead to a speaker-level granularity in which to analyse conversational interactions. Such pipeline of analysis itself could also be useful outside of the context of therapy, possibly to gain further insight into communication as a whole. Additionally, issues of such automatic transcription pipeline are highlighted, providing hopefully useful knowledge to future research.

While the prediction of working alliance may not yet be fruitful, certain aspects of the presented methodology may already be useful to therapists. For example, the visualisation of affect over time may provide insights to therapists who can use it to make judgements about their patients and sessions without needing a fully automatic prediction of working alliance. Also, the custom RobBERT model's ability to capture affect from Dutch text may provide a useful tool or inspiration for future research.

A. Appendices

A.1 WAI-S

1. Een resultaat van deze sessies is dat het voor mij duidelijker is hoe ik zou kunnen veranderen.

   ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

2. Wat ik doe in therapie, geeft mij een nieuwe kijk op mijn probleem.

   ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

3. Ik geloof dat mijn therapeut(e) mij aardig vindt.

   ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

4. Mijn therapeut(e) en ikzelf werken samen bij het bepalen van de doelstellingen voor mijn therapie.

   ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

5. Mijn therapeut(e) en ik respecteren elkaar.

   ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

6. Mijn therapeut(e) en ik werken naar de doelstellingen toe die we beiden goedkeurden.

   ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

7. Ik voel dat mijn therapeut(e) mij apprecieert.

   ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

8. Wij zijn het eens over wat voor mij belangrijk is om aan te werken

   ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

9. Ik voel dat mijn therapeut(e) om mij geeft, zelfs wanneer ik dingen doe die hij/zij niet goedkeurt.

   ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

10. Ik voel dat de dingen die ik in therapie doe, mij zullen helpen om de veranderingen die ik wil, te bereiken.

    ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

11. We hebben ons een goed begrip gevormd van het soort veranderingen die goed zouden zijn voor mij.

    ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

12. Ik geloof dat de manier waarop we aan mijn probleem werken, de juiste is.

    ZELDEN OF NOOIT / SOMS / DIKWIJLS / HEEL VAAK / ALTIJD

Figure 7. Patient version of the WAI-S, the observer version is identical except for using a third person perspective, Figure by Vollebregt (2023)

A.2 WAI-SRT

Working Alliance Inventory – Short Revised - Therapist (WAI-SRT)

Instructions: Below is a list of statements about experiences people might have with their client. Some items refer directly to your client with an underlined space – as you read the sentences, mentally insert the name of your client in place of ___ in the text.

IMPORTANT!!! Please take your time to consider each question carefully.

1. ___ and I agree about the steps to be taken to improve his/her situation.

①	②	③	④	⑤
Seldom	Sometimes	Fairly Often	Very Often	Always

2. I am genuinely concerned for ___'s welfare.

⑤	④	③	②	①
Always	Very Often	Fairly Often	Sometimes	Seldom

3. We are working towards mutually agreed upon goals.

①	②	③	④	⑤
Seldom	Sometimes	Fairly Often	Very Often	Always

4. ___ and I both feel confident about the usefulness of our current activity in therapy.

①	②	③	④	⑤
Seldom	Sometimes	Fairly Often	Very Often	Always

5. I appreciate ___ as a person.

⑤	④	③	②	①
Always	Very Often	Fairly Often	Sometimes	Seldom

6. We have established a good understanding of the kind of changes that would be good for ___.

⑤	④	③	②	①
Always	Very Often	Fairly Often	Sometimes	Seldom

7. ___ and I respect each other.

①	②	③	④	⑤
Seldom	Sometimes	Fairly Often	Very Often	Always

8. ___ and I have a common perception of his/her goals.

⑤	④	③	②	①
Always	Very Often	Fairly Often	Sometimes	Seldom

9. I respect ___ even when he/she does things that I do not approve of.

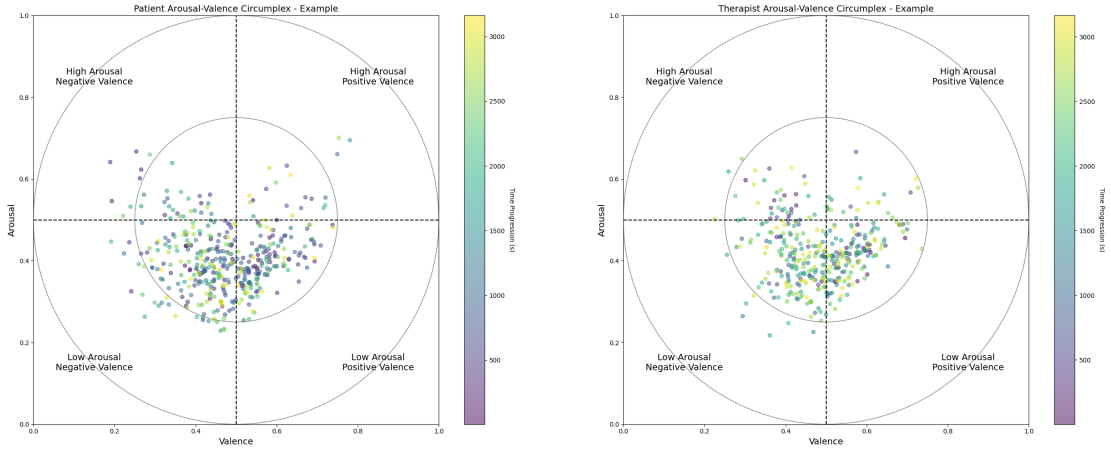
①	②	③	④	⑤
Seldom	Sometimes	Fairly Often	Very Often	Always

10. We agree on what is important for ___ to work on.

⑤	④	③	②	①
Always	Very Often	Fairly Often	Sometimes	Seldom

Figure 8. Therapist version of the WAI-SRT. Figure associated with Horvath and Greenberg (1989)

A.3 Affect visualisations



(a) Therapist (b) Patient

Figure 9. Affect circumplex models over time for therapist and patient

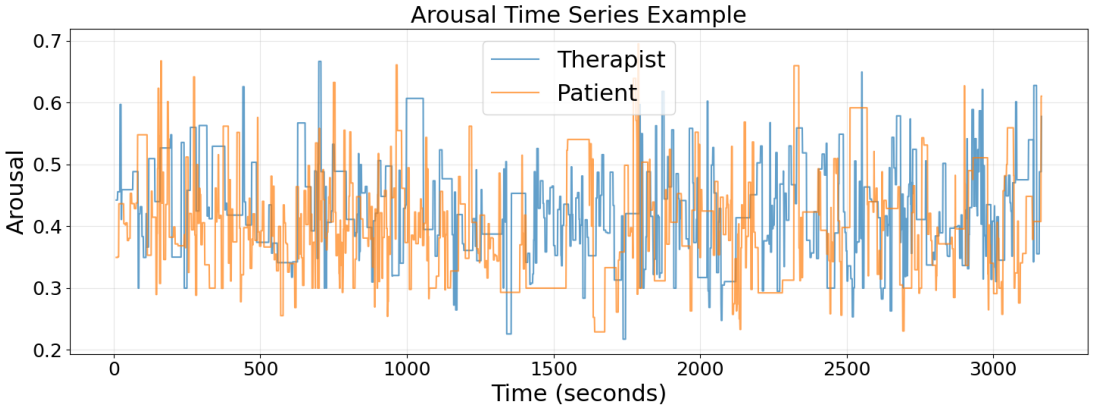


Figure 10. Example of Arousal as resampled time-series

A.4 Final dataset distributions

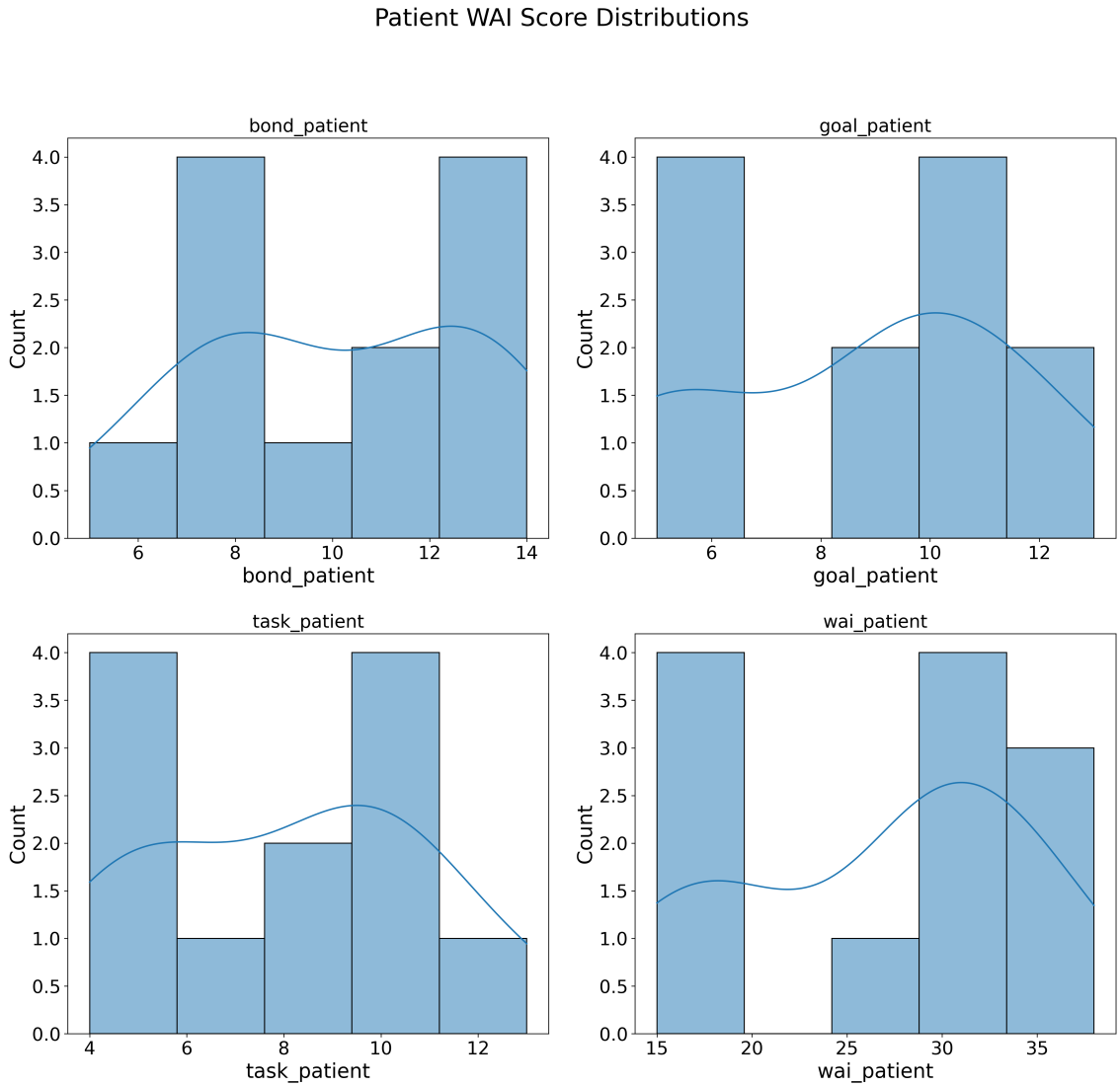


Figure 11. WAI patient score distributions in final dataset

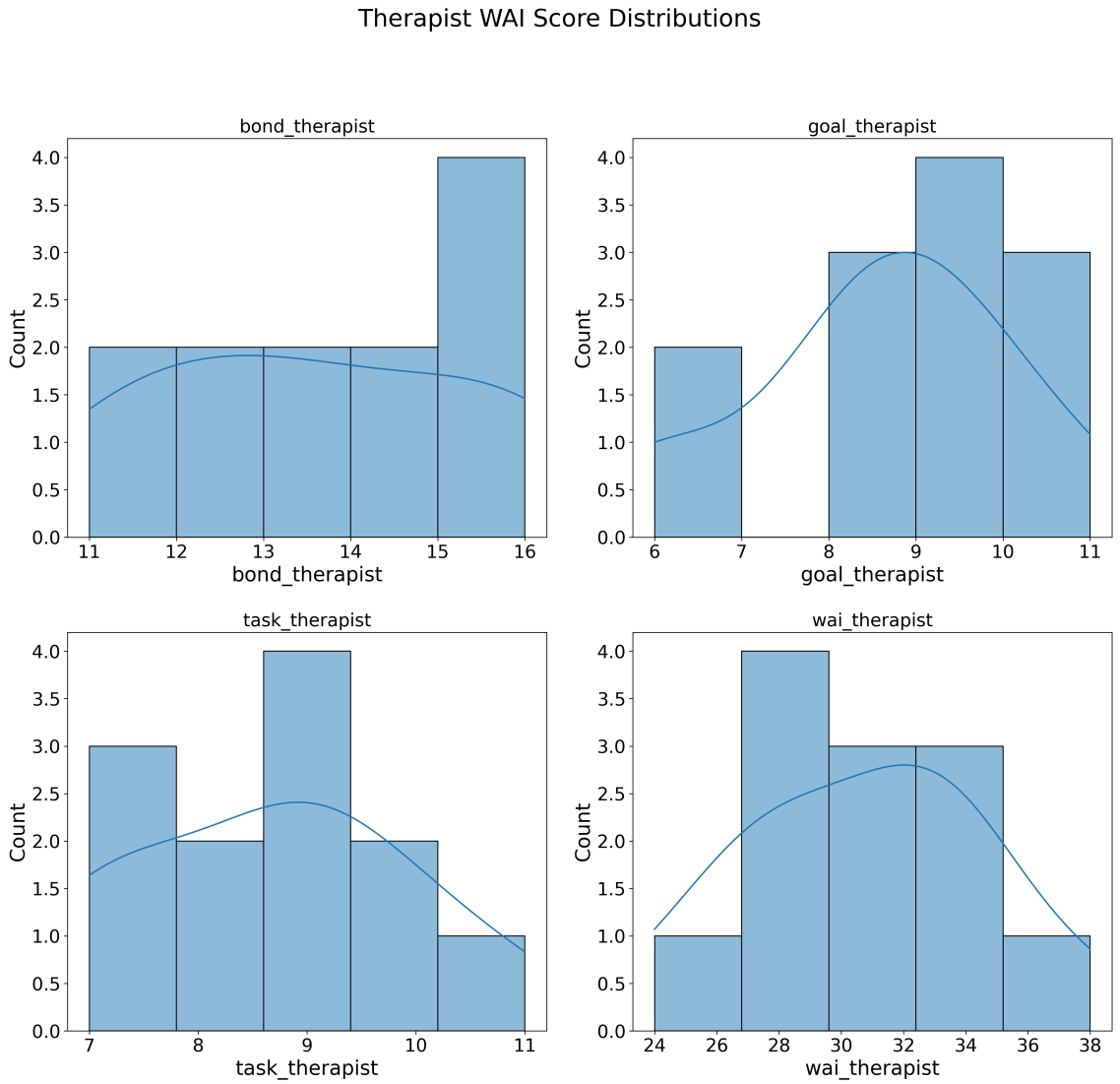


Figure 12. WAI therapist score distributions in final dataset

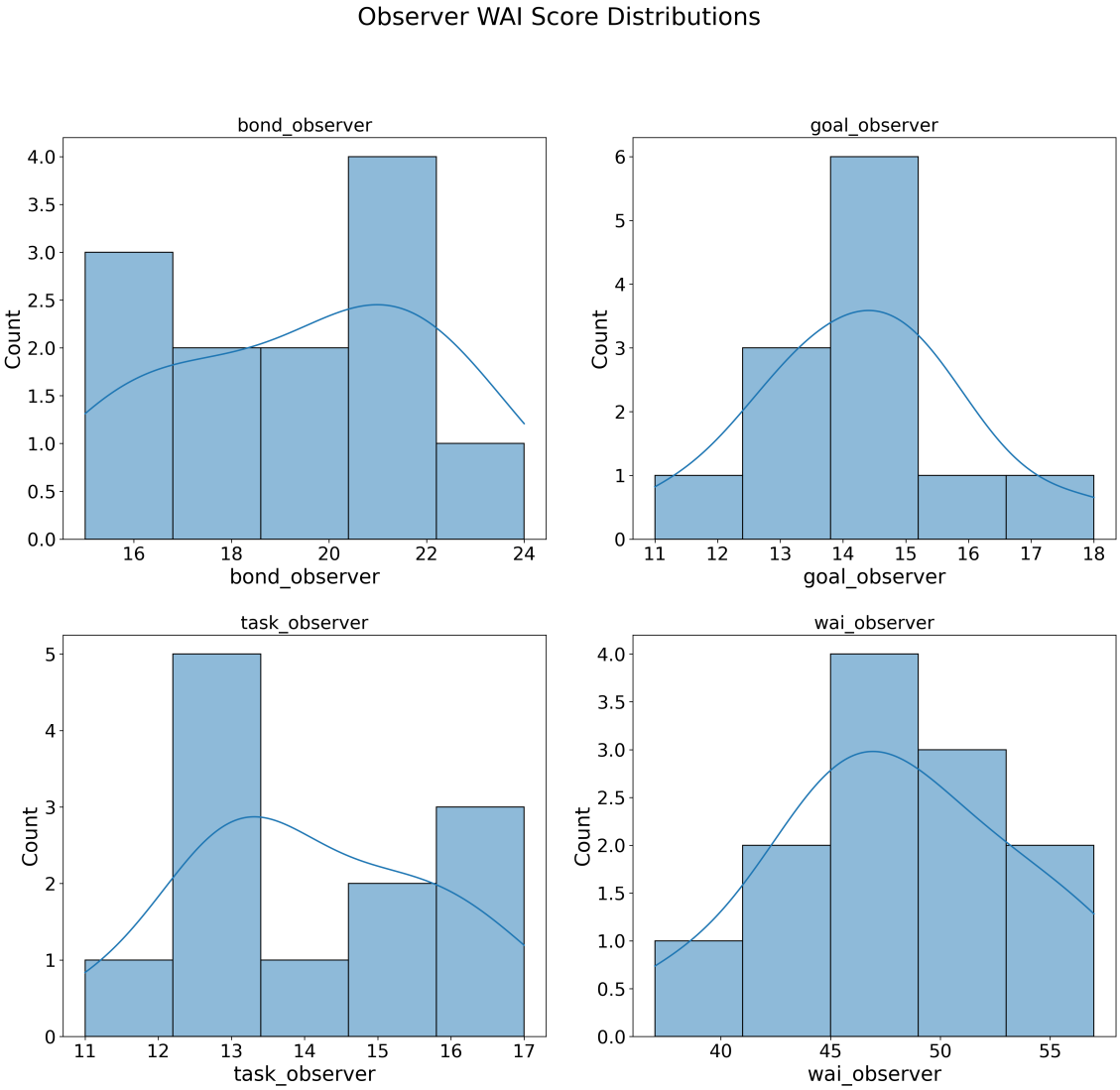


Figure 13. WAI observer score distributions in final dataset

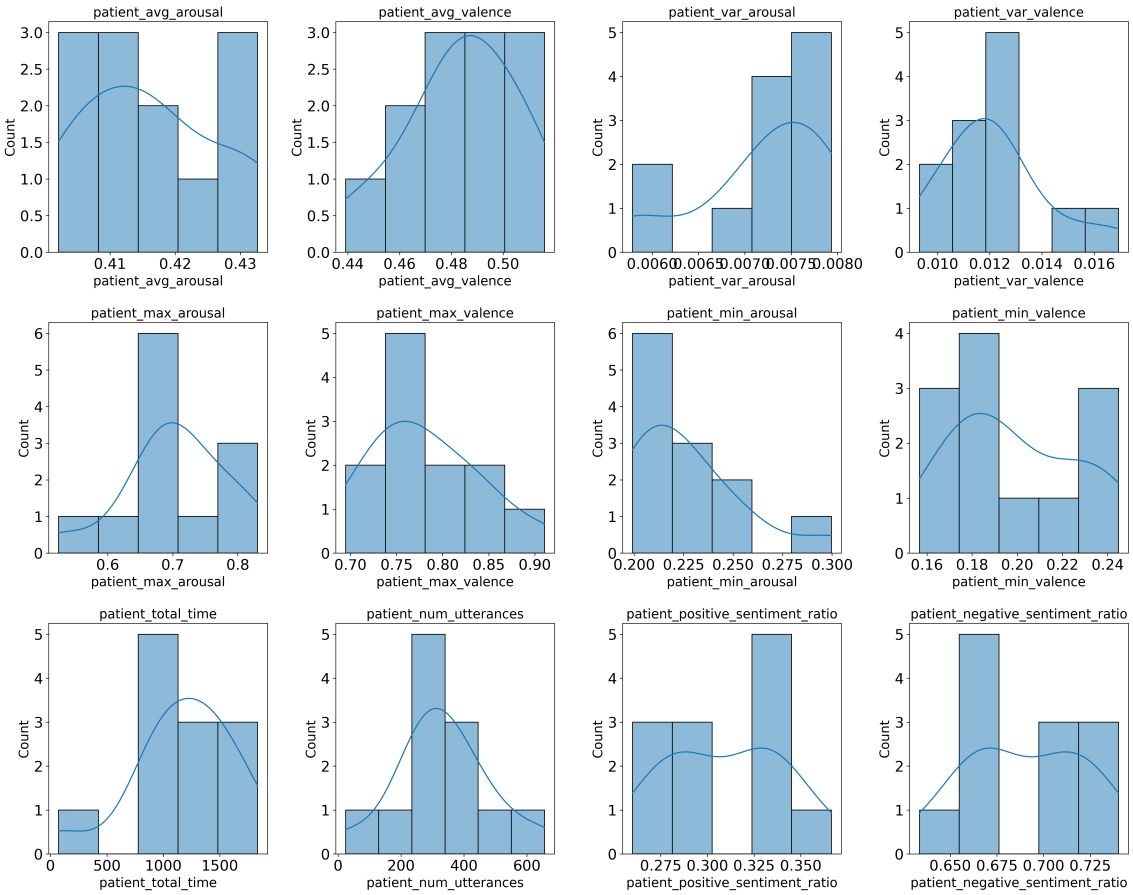


Figure 14. Patient related feature distributions in final dataset

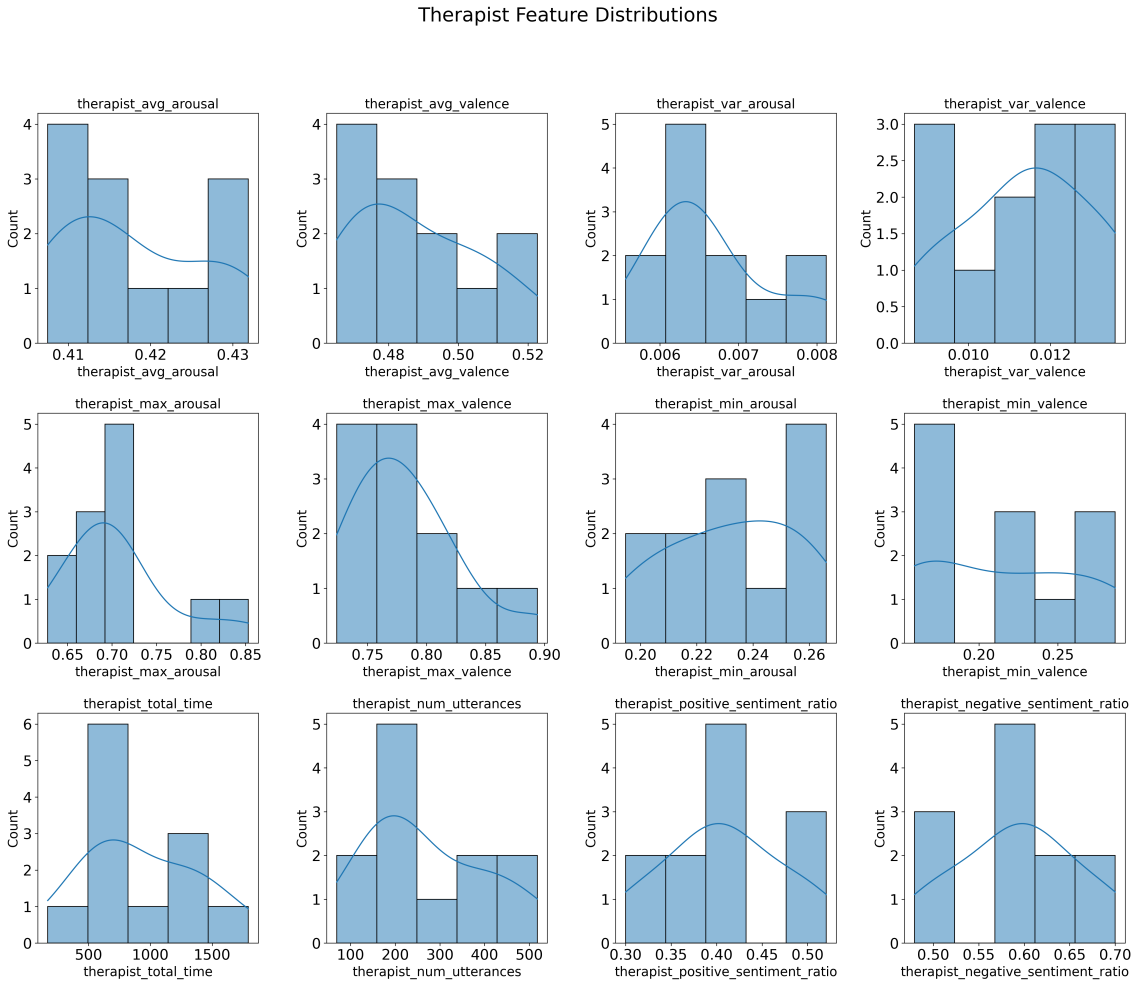


Figure 15. Therapist related feature distributions in final dataset

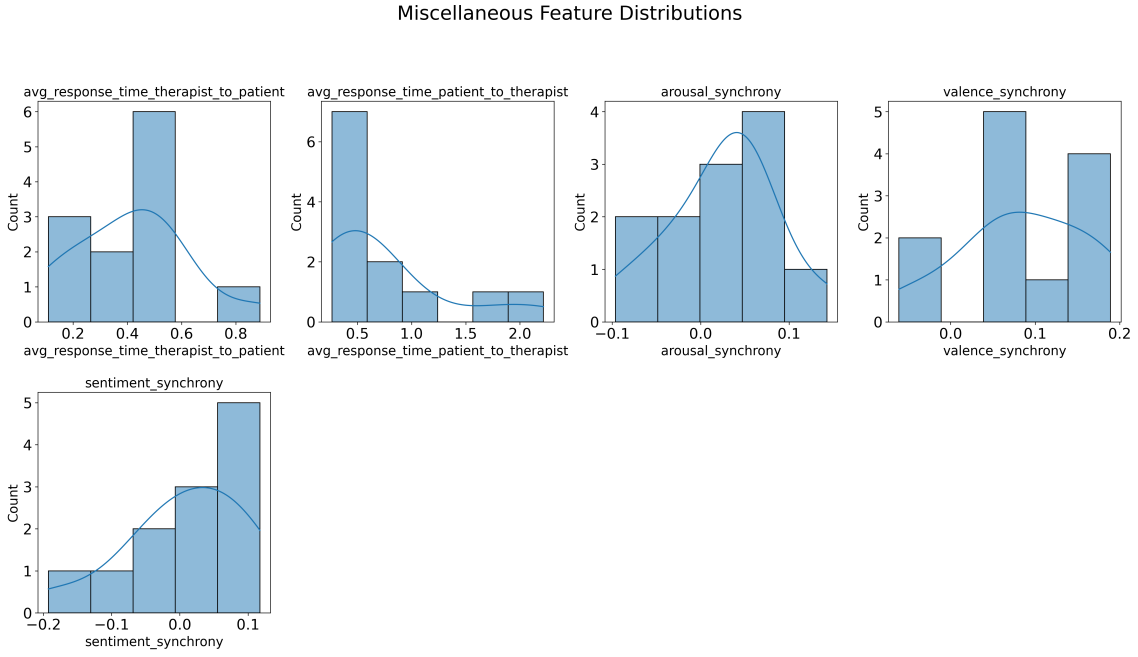


Figure 16. Miscellaneous feature distributions in final dataset

A.5 KNN regression performance

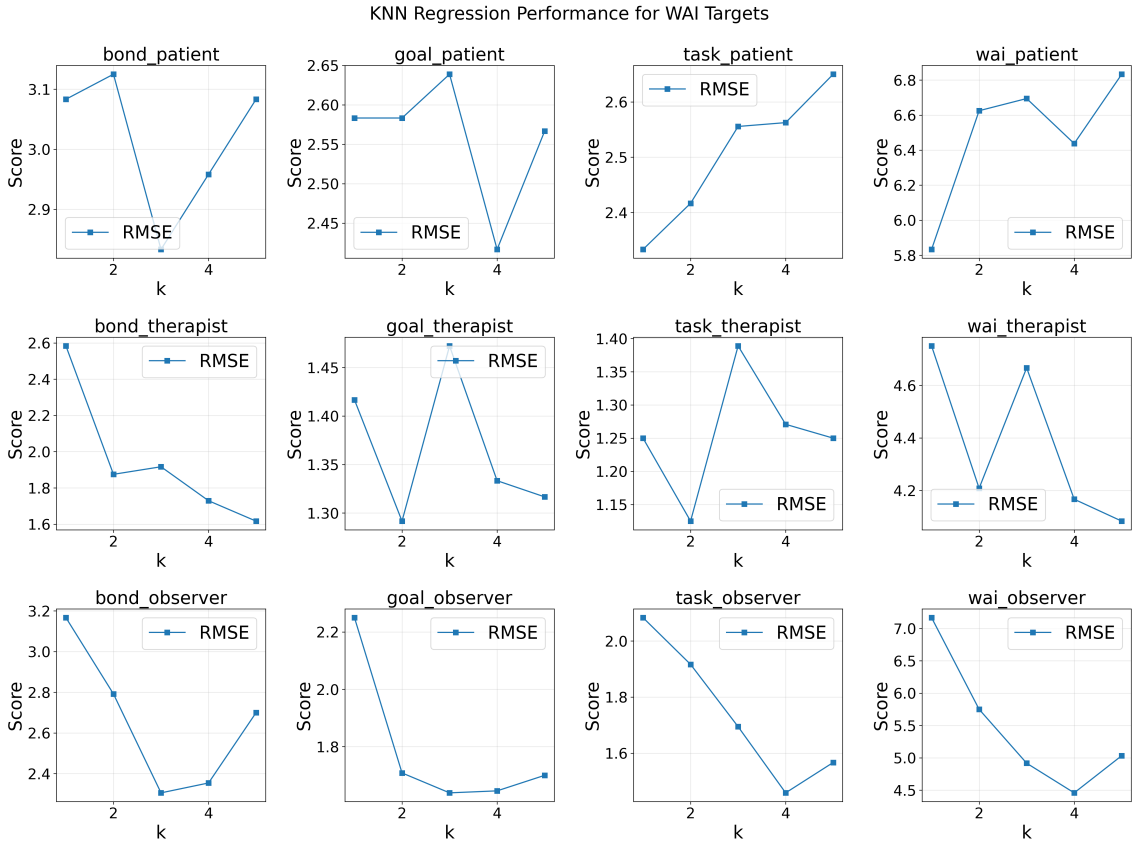


Figure 17. RMSE of KNN regression for different K values

Bibliography

Akesbi, K. and Gandhi, S. (2024). "diarizers: A repository for fine-tuning speaker diarization models". `https://github.com/huggingface/diarizers`.

Anderson, L. A. and Dedrick, R. F. (1990). Development of the trust in physician scale: a measure to assess interpersonal trust in patient-physician relationships. *Psychological reports*, 67(3_suppl):1091–1100.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Bain, M., Huh, J., Han, T., and Zisserman, A. (2023). Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Bayerl, S. P., Roccabruna, G., Chowdhury, S. A., Ciulli, T., Danieli, M., Riedhammer, K., and Riccardi, G. (2022). What can speech and language tell us about the working alliance in psychotherapy. *arXiv preprint arXiv:2206.08835*.

Beck, A. T. (1979). *Cognitive therapy of depression*. Guilford press.

Berry, D. S. and Pennebaker, J. W. (1993). Nonverbal and verbal emotional expression and health. *Psychotherapy and psychosomatics*, 59(1):11–19.

Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, research & practice*, 16(3):252.

Bradley, M. M. and Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59.

Bradley, M. M. and Lang, P. J. (1999). Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology ....

Branigan, H. P., Pickering, M. J., McLean, J. F., and Cleland, A. A. (2007). Syntactic alignment and participant role in dialogue. *Cognition*, 104(2):163–197.

Bredin, H. and Laurent, A. (2021). End-to-end speaker segmentation for overlap-aware resegmentation. *arXiv preprint arXiv:2104.04045*.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., and Gill, M.-P. (2020). Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7124–7128. IEEE.

Brennan, S. E. and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22(6):1482.

Bronstein, I., Nelson, N., Livnat, Z., and Ben-Ari, R. (2012). Rapport in negotiation: The contribution of the verbal channel. *Journal of Conflict Resolution*, 56(6):1089–1115.

Bruijniks, S. J., Lemmens, L. H., Hollon, S. D., Peeters, F. P., Cuijpers, P., Arntz, A., Dingemanse, P., Willems, L., Van Oppen, P., Twisk, J. W., et al. (2020). The effects of once-versus twice-weekly sessions on psychotherapy outcomes in depressed patients. *The British Journal of Psychiatry*, 216(4):222–230.

Busseri, M. A. and Tyler, J. D. (2003). Interchangeability of the working alliance inventory and working alliance inventory, short form. *Psychological assessment*, 15(2):193.

Canavan, A., Graff, D., and Zipperlen, G. (1997a). Callhome american english speech. Linguistic Data Consortium. LDC97S42.

Canavan, A., Graff, D., and Zipperlen, G. (1997b). Callhome german speech. Linguistic Data Consortium. LDC97S43.

Castonguay, L. G., Constantino, M. J., and Holtforth, M. G. (2006). The working alliance: Where are we and where should we go? *Psychotherapy: Theory, Research, Practice, Training*, 43(3):271.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

De Bruyne, L., De Clercq, O., and Hoste, V. (2021). Prospects for dutch emotion detection: Insights from the new emotionl dataset. *Computational Linguistics in the Netherlands Journal*, 11:231–255.

Delobelle, P., Winters, T., and Berendt, B. (2020). Robbert: a dutch roberta-based language model. *arXiv preprint arXiv:2001.06286*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Duran, N. D., Paxton, A., and Fusaroli, R. (2019). Align: Analyzing linguistic interactions with generalizable techniques—a python library. *Psychological methods*, 24(4):419.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Ekman, P. and Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Fuertes, J. N., Toporovsky, A., Reyes, M., and Osborne, J. B. (2017). The physician-patient working alliance: Theory, research, and future possibilities. *Patient Education and Counseling*, 100(4):610–615.

Hatcher, R. L. and Gillaspy, J. A. (2006). Development and validation of a revised short version of the working alliance inventory. *Psychotherapy research*, 16(1):12–25.

Horvath, A. O., Del Re, A., Flückiger, C., and Symonds, D. (2011). Alliance in individual psychotherapy. *Psychotherapy*, 48(1):9.

Horvath, A. O. and Greenberg, L. S. (1989). Development and validation of the working alliance inventory. *Journal of counseling psychology*, 36(2):223.

Horvath, A. O. and Symonds, B. D. (1991). Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of counseling psychology*, 38(2):139.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Kamath, R., Ghoshal, A., Eswaran, S., and Honnavalli, P. (2022). An enhanced context-based emotion detection model using roberta. In *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pages 1–6. IEEE.

Kamilakesbi (2024a). speaker-segmentation-fine-tuned-callhome-deu. `https://huggingface.co/diarizers-community/speaker-segmentation-fine-tuned-callhome-deu`. Accessed: 12 May 2024.

Kamilakesbi (2024b). speaker-segmentation-fine-tuned-callhome-eng. https://huggingface.co/diarizers-community/speaker-segmentation-fine-tuned-callhome-eng. Accessed: 12 May 2024.

Kashdan, T. B. and Rottenberg, J. (2010). Psychological flexibility as a fundamental aspect of health. *Clinical psychology review*, 30(7):865–878.

Klerman, G. L., Weissman, M. M., Rounsaville, B., and Chevron, E. S. (1996). Interpersonal psychotherapy for depression.

Loria, S. et al. (2018). textblob documentation. *Release 0.15*, 2(8):269.

Mauksch, L. B., Dugdale, D. C., Dodson, S., and Epstein, R. (2008). Relationship, communication, and efficiency in the medical encounter: creating a clinical model from a literature review. *Archives of internal medicine*, 168(13):1387–1395.

Müller-Frommeyer, L. C., Frommeyer, N. A., and Kauffeld, S. (2019). Introducing rlsm: An integrated metric assessing temporal reciprocity in language style matching. *Behavior Research Methods*, 51:1343–1359.

Negri, A., Christian, C., Mariani, R., Belotti, L., Andreoli, G., and Danskin, K. (2019). Linguistic features of the therapeutic alliance in the first session: a psychotherapy process study. *Research in Psychotherapy: Psychopathology, Process, and Outcome*, 22(1).

Newhill, C. E., Safran, J. D., and Muran, J. C. (2003). *Negotiating the therapeutic alliance: A relational treatment guide*. Guilford Press.

OpenAI (2024). Chatgpt-4.0. https://www.openai.com. Accessed: June 18, 2024.

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., and Narayanan, S. (2022a). A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, 72:101317.

Park, T. J., Koluguri, N. R., Balam, J., and Ginsburg, B. (2022b). Multi-scale speaker diarization with dynamic scale weighting. *arXiv preprint arXiv:2203.15974*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71.

Pinto, R. Z., Ferreira, M. L., Oliveira, V. C., Franco, M. R., Adams, R., Maher, C. G., and Ferreira, P. H. (2012). Patient-centred communication is associated with positive therapeutic alliance: a systematic review. *Journal of physiotherapy*, 58(2):77–87.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R. D., and Bengio, Y. (2021). SpeechBrain: A general-purpose speech toolkit. arXiv:2106.04624.

Reddy, C. K., Gopal, V., and Cutler, R. (2022). Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 886–890. IEEE.

Rogers, D. T. (2015). Further validation of the learning alliance inventory: The roles of working alliance, rapport, and immediacy in student learning. *Teaching of Psychology*, 42(1):19–25.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Sainburg, T. (2019). timsainb/noisereduce: v1.0.

Sharma, E. and De Choudhury, M. (2018). Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13.

Speed, L. J. and Brysbaert, M. (2023). Ratings of valence, arousal, happiness, anger, fear, sadness, disgust, and surprise for 24,000 dutch words. *Behavior Research Methods*, pages 1–17.

Templeton, E. M., Chang, L. J., Reynolds, E. A., Cone LeBeaumont, M. D., and Wheatley, T. (2022). Fast response times signal social connection in conversation. *Proceedings of the National Academy of Sciences*, 119(4):e2116915119.

Tracey, T. J. and Kokotovic, A. M. (1989). Factor structure of the working alliance inventory. *Psychological Assessment: A journal of consulting and clinical psychology*, 1(3):207.

Vail, A., Girard, J., Bylsma, L., Cohn, J., Fournier, J., Swartz, H., and Morency, L.-P. (2022). Toward causal understanding of therapist-client relationships: A study of language modality and social entrainment. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 487–494.

Vail, A. K. (2023). *Multimodal and Social Modeling of Client-Therapist Interaction*. PhD thesis, Carnegie Mellon University Pittsburgh, PA.

Vail, A. K., Liebson, E., Baker, J. T., and Morency, L.-P. (2018). Toward objective, multifaceted characterization of psychotic disorders: lexical, structural, and disfluency markers of spoken language. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 170–178.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vollebregt, R. (2023). A multimodal approach to working alliance detection in therapist-patient psychotherapy using deep learning models. Master's thesis.

Woodard, J. and Nelson, J. (1982). An information theoretic measure of speech recognition performance. In *Workshop on standardisation for speech I/O technology, Naval Air Development Center, Warminster, PA*.