THE APPLICATION OF SENTIMENT ANALYSIS TO A PSYCHOTHERAPY SESSION: AN EXPLORATORY STUDY USING FOUR GENERAL-PURPOSE LEXICONS

GONÇALO MARMELO DA SILVA VEIGA

Orientador de Dissertação:
PROF. DOUTOR ANTÓNIO PAZO PIRES

Coordenador de Seminário de Dissertação:
PROF. DOUTOR ANTÓNIO PAZO PIRES

Tese submetida como requisito parcial para a obtenção do grau de:

MESTRE EM PSICOLOGIA

Especialidade em Psicologia Clínica

2019

Dissertação de Mestrado realizada sob a orientação de Prof. Doutor António Pazo Pires, apresentada no ISPA – Instituto Universitário para obtenção de grau de Mestre na especialidade de Psicologia Clínica.

Agradecimentos

Ao Professor Doutor António Pazo Pires, o meu orientador de dissertação, pelo seu apoio ao longo do ano, pela sua paciência e generosidade com o seu tempo, por me permitir que eu fizesse esta pesquisa na interceção da psicologia com a linguística computacional. E, claro, por me deixar sentar na fila da frente nas suas aulas, estando eu mesmo a um oceano de distância.

Aos meus pais, António e Ilda, pelo seu amor e apoio incondicional.

Ao meu irmão, João, pelas longas conversas.

A Analise, minha esposa. Por tudo.

Table of Contents

Table of Figures

Abstract

In this study we explore the application of sentiment analysis to a complete and in-person psychotherapy session. Sentiment analysis is a text mining technique that allows for the analysis, interpretation, and visualization of textual data. We investigate how we can apply a lexicon-based approach to analyze clinical session data, using four general-purpose lexicons available within an open-source statistical programming language environment, R.

We conducted our study by comparing the performance of four general-purpose lexicons to the performance of n = 52 human raters, using inter-rater reliability (IRR) and intraclass correlation (ICC) measurements. Our findings suggest there is low to moderate agreement between human ratings and lexicon generated ratings, depending on the lexicon used. There are some benefits in applying a lexicon-based sentiment analysis approach to psychotherapy session data, namely the way it efficiently processes and analyses data and allows for novel visualizations of psychotherapy data. We recommend further investigation into the application of sentiment analysis as a technique, focusing on the performance of specific-purpose lexicons. We also recommend further research into comparing the performance of lexicon-based approaches to text classification approaches to the analysis of psychotherapy data.

Keywords: sentiment analysis, lexicon-based approach, therapy session data.

Resumo

Neste estudo exploramos a aplicação da análise de sentimento a uma sessão integral de psicoterapia. A análise de sentimento é uma técnica que se enquadra dentro da mineração de texto e que permite a análise, interpretação e visualização de dados textuais. Nesse sentido, investigamos como poderemos aplicar uma abordagem assente no uso de dicionários de termos para analisar dados de sessões clínicas, com o recurso a quatro dicionários de termos gerais. Para o efeito, utilizamos uma aplicação de programação e de análise estatística, de acesso livre, chamada R.

O estudo comparou a performance de quatro dicionários de termos gerais à performance de n = 52 avaliadores humanos, recorrendo à análise da concordância entre avaliadores (IRR) e correlação intraclasses (ICC). A nossa pesquisa sugere que existe uma baixa a moderada concordância entre as cotações de avaliadores humanos e as cotações geradas pelos dicionários de termos, dependendo dos dicionários utilizados. Existem alguns claros benefícios na aplicação de uma análise de sentimento baseada em dicionários de termos a dados de sessões de psicoterapia, nomeadamente, a forma rápida e eficiente como processa, analisa e permitir visualizar os dados de sessões de psicoterapia. Recomendamos o aprofundamento da exploração da aplicação da análise de sentimento a sessões clínicas, com o recurso a dicionários de termos específicos a psicoterapia. Recomendamos também uma aposta na pesquisa focada na comparação de análise de sentimento baseadas em dicionários de termos com a análise de sentimento baseada na classificação de texto na análise de dados clínicos.

Palavras-chave: análise de sentimento, dicionário de termos, dados clínicos de psicoterapia.

1. Introduction

There are a number of technological advances that have had a tremendous impact on the development of science (Brooks, 1994). Several of these technological innovations have cultivated whole new fields of research across the sciences, within physics, astronomy, genetics, neuroscience, as well as in the social sciences. Often coupled with these technological advances are methodological breakthroughs that propel us forward in the way we manufacture scientific work and generate knowledge. In the past century alone, we have witnessed contributions both from academia and industry that have allowed us to progress in the way we make observations, take measurements, and analyze data (Imel, Steyvers, & Atkins, 2015; Iliev, Dehghani, & Sagi, 2015). In fact, technology has been "a source of otherwise unavailable instrumentation and techniques needed to address novel and more difficult scientific questions more efficiently" (Brooks, 1994, p. 477).

Recent developments in computer science and the proliferation of human-generated data open the door to cutting-edge analyses of psychological data and are expected to have a tremendous impact in psychology and psychotherapy (Boyd & Pennebaker, 2016; Montag, Duke, & Markowetz, 2016). Some authors even go as far as to suggest that "psychotherapy is at the verge of a technological-inspired revolution" (Imel, Caperton, Tanana, & Atkins, 2017, p. 1). It might not be an overstatement. Traditionally, psychological research methods have remained largely unchanged. We have been relying on laboratory experiments, where subjects are prodded and expected to behave as they would in the real world; on self-reporting, via surveys and questionnaires, applied to patients who might answer them in a socially desirable manner rather than truthfully; and on human coders, who require special training and considerable time and cost (Pace, et al., 2016; Boyd R. L., 2017). By borrowing some of the techniques and methodologies emanating from computer science, computational linguistics, and other related fields, psychotherapy might be positioned to: make advances in identifying the underlying processes and mechanisms related to how psychotherapy works and helps clients change; evaluate and predict patients' responses to psychotherapy and therapists' input; and provide performance-based feedback to therapists relying on actual session data and not on delayed nor biased self-reporting (Imel, Steyvers, & Atkins, 2015; Imel, Caperton, Tanana, & Atkins, 2017; Hirsch, et al., 2018). In

addition, these techniques, methods and analyses also have the potential to be integrated into technologies that become part of the daily lives of patients (Holmes, et al., 2018).

In the past, access to sensors, recording devices, powerful computers, and statistical software was generally only afforded to those affiliated with affluent academic institutions or private corporations. In contrast, today a wide portion of mental health care professionals have access to tools that enable them to generate and analyze their own data (Owen & Imel, 2016). To illustrate, many recent cell phone models are equipped with capable microphones that can competently capture data from therapy or counselling sessions and store it in the devices or save it to a variety of cloud storage services (Harari, et al., 2016). Speech-to-text transcription software, although imperfect, offers a cheap and fast way to convert session audio to text (Ziman, Heusser, Fitzpatrick, Field, & Manning, 2018). Therapists' notes and patients' notebooks or drawings can be easily digitized with non-specialized cameras. Optical character recognition (OCR) software has the ability to faithfully convert typed or handwritten text to machine-readable textual data better than ever (Reshma, James, Kavya, & Saravanan, 2016; Dalianis, 2018). Modern personal computers have become extremely efficient at performing sophisticated operations, often only limited by their own storage capacity, which can be expanded at a reasonable cost. Moreover, proprietary statistical software have now suitable competitors in the free open-source space allowing more individuals to use statistical programming software to process, analyze, and visualize their own data (Yarkoni, 2012).

In order to make use of all this data, psychotherapists and psychology researchers can appropriate methods from the nascent field of data science, a field of applied research at the interface of computer science, software engineering, and statistics. In its arsenal of approaches to analyzing data, data science utilizes machine learning, deep learning, artificial intelligence, natural language processing, and text mining to generate insight from data. In this article we will be focusing on sentiment analysis, one of text mining's techniques, to analyze psychological data.

Sentiment analysis, otherwise known as opinion mining or subjectivity analysis, allows us to extract subjectivity from text by 1) evaluating discrete or overall polarity (if a word or any sized corpus is positive, neutral or negative) and its intensity or strength (by rating how positive and negative it is on a scale), and by 2) identifying emotion (Pang & Lee, 2008; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011; Liu & Zhang, 2012). This technique uses two approaches: a lexicon-based approach, which requires the use of established dictionaries, where lists of human-

annotated words are used to match those found in the text subject to analysis; and a text classification approach, which requires building classifiers from labeled instances of text using machine-learning methods (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011).

While the analysis of psychological text has a long and rich history (Boyd R. L., 2017) and the application of computer software to the analysis of psychological text is short of being novel (Tausczik & Pennebaker, 2010; Boyd & Pennebaker, 2016), the use of sentiment analysis as a tool to measure a patient's emotional status intra-session and in the course of treatment is open for exploration. The tracking of sentiment and emotion allows therapists to follow patients process of change across one or multiple sessions, evaluate responses to therapists' interventions, measure the words associated with particular entities or situations, and provide the patient a session summary, highlighting key moments and topics discussed (Imel, Caperton, Tanana, & Atkins, 2017).

The overall goal of this article is to explore the value and competence of computational analyses of therapy session data, using a free and widely available statistical programming software, R. In order to evaluate the application of a dictionary-based approach of sentiment analysis to a full therapy session, we compare the performance of several dictionaries against the ratings generated by human raters.

2. Background and Related Research

A Brief History of Computerized Text Analysis

The analysis of words, speech, and narrative in a clinical setting can be traced all the way back to the beginning of modern psychology and the birth of psychoanalysis. At the turn of the 20th century, Sigmund Freud used language as a means to access the unconscious, understand the functioning of the mind, and as a therapeutic device. His analyses of slips-of-the-tongue, forgetfulness, free association, and dreams relied on a careful and laborious examination of words (Freud, 1900/1913; 1914/1990; 1973/1991). In the 1920s and '30s, others followed with the development of projective tests, like Hermann Rorschach's ink blot test and Henry Murray and Christiana Morgan's Thematic Apperception Test, evaluating personality structure and functioning, and interpersonal relationships through the analysis of the narratives of test takers. In the 1950s, Gottschalk and colleagues developed a content analysis technique that tried to identify Freudian themes in subjects speech. Subjects were encouraged to free-associate to a voice recorder for five minutes and transcribed text samples would then be broken down and analyzed by judges (Tausczik & Pennebaker, 2010).

In the following decades, many other approaches to text analysis were developed, including concordance analysis, conversation analysis, qualitative text analysis, discourse analysis, linguistic content analysis, and network analysis (Melina, 1997). These qualitative approaches are rich in content and in interpretative value but possess serious drawbacks.

First, scaling-up is a challenge. If an hour-long session takes ten hours to be transcribed and analyzed, ten sessions take ten times as long. The fact that these analyses depend exclusively on the work of human raters and coders makes them impractical as the amount of textual data increases (Iliev, Dehghani, & Sagi, 2015; Imel, Steyvers, & Atkins, 2015). Second, even with thorough training, achieving high levels of inter-rater agreement is difficult and costly (Tausczik & Pennebaker, 2010; Pace, et al., 2016). Third, in some particular instances, human raters report negative changes in humor when rating depressing text (Tausczik & Pennebaker, 2010).

Computerized text analysis in psychology first appeared in the 1960s with the Harvard General Inquiry, by Phillip Stone and his colleagues (Tausczik & Pennebaker, 2010; Boyd & Pennebaker, 2016). The General Inquiry was able to successfully identify and distinguish mental

disorders from text, assess personality dimensions, and evaluate speech (Tausczik & Pennebaker, 2010, p. 26). However, the General Inquiry and others based on it had two major problems. One the one hand, the dictionaries that were built into the software reflected the highly specialized theoretical interests of the researchers that created them. This made them unreliable for use both with general corpora and in specific use-cases outside the scope of the dictionaries. As Boyd (2017, p. 164) put it, "a system of analyzing texts for a specific Neo-Freudian process had little use when researching [Carl Jung's] extroversion." On the other hand, the dictionaries and algorithms used in these early computational text analysis programs were opaque and hard to interpret. They did not allow users insight into what language variables were being manipulated, what weights were being applied to words, nor access to the calculations (Tausczik & Pennebaker, 2010, p. 26).

In 1999 James W. Pennebaker and colleagues introduced a software application called Linguistic Inquiry and Word Count (LIWC) designed "to provide an efficient and effective method for studying the various emotional, cognitive, structural, and process components present in individuals' verbal and written speech samples" (Pennebaker, Francis, & Booth, 2001, p. 12). Their innovation was based on the realization that, throughout most of the history of language analysis in psychology, studies focused mostly on linguistic content rather than on linguistic style or form (Boyd & Pennebaker, 2016). In order to integrate both content and style to their methodological approach in computerized text analysis, the software and its dictionary were developed to reflect these two types of approaches to analyzing data. Hence, LIWC processes text by mapping words into two different domains: it categorizes words based on common content, such as affiliation, power, religion, money, emotions, or biological processes, and is also able to identify function words, such as pronouns, verbs, articles, and adjectives (Boyd & Pennebaker, 2016). It also takes into account word length, word count, and tense (Iliev, Dehghani, & Sagi, 2015).[1]

LIWC's ease of use out-of-the-box, along with the statistical and psychometric validation using external psychological data has enabled researchers to analyze text in a fast and reliable manner. Its focus on linguistic style and on function words in particular has allowed to uncover many interesting patterns. For instance, the use of first-person singular pronouns has been found to be associated with negative experiences, depression, and people in lower positions of power (Boyd R. L., 2017). Newman and colleagues (Newman, Pennebaker, Berry, & Richards, 2003)

[1] For a full account of the categories and dimensions, see Boyd & Pennebaker (2016).

identified that the use of particular function words allow researchers to distinguish truthful statements from deceptions or lies with at least a 61% success rate (Newman, Pennebaker, Berry, & Richards, 2003). Also using LIWC, Al-Mosaiwi and Johnstone (2018) found that people who post in anxiety, depression, and suicide ideation-oriented online forums use more absolutist words than those in control forums.

Applications in Clinical Psychology and Psychotherapy

The application of machine learning, natural language processing, and text mining techniques can radically transform the practice, teaching, training, and research in psychotherapy. This is especially pertinent when we realize there is still much to learn and figure out about process, how or why psychotherapy works, and which factors play into the development of therapists' competency over time (Schroder, Orlinsky, Ronnestad, & Willutzki, 2015; Goldberg, et al., 2016). While some psychologists might wall-up defensively, wary they will have to give up their methods and techniques, we have to assuage those concerns and remind them these techniques are meant to augment, not diminish nor replace, their skills and learning opportunities (Montag, Duke, & Markowetz, 2016).

The need for therapist evaluation, reliable feedback from patients, and data-based continuous supervision are critical to the improvement of skills and competency in psychotherapists. Several automated feedback systems tailored to psychotherapy have surfaced in the past five years (Xiao, Imel, Georgiou, Atkins, & Narayanan, 2015; Can, et al., 2016; Hirsch, et al., 2018) and the responses from therapists to these systems seem encouraging. Hirsch and colleagues (Hirsch, et al., 2018) investigated psychotherapists responses to CORE-MI, an automated feedback system for counselors applying Motivational Interviewing in therapy sessions. This system is designed to give counselors feedback based on session data in the form a report card, with a visual summary of the counselling session, including percentages of therapist and client talk time, talk turns, the quality of the counselor reflections, among others. It offers insight into a therapist's overall and specific competency, levels of empathy, and adherence to the specific therapy theoretical principles. The researchers found that therapists were overall receptive to the automated feedback and were confident in the accuracy of CORE-MI measures, even though they did not understand how they were derived (Hirsch, et al., 2018, p. 10).

It is also well documented that psychotherapy novices and trainees might experience acute performance anxiety as they start out (Skovholt & Ronnestad, 2003). New simulation systems offer an opportunity for trainees to practice with computerized conversational agents in a safe testing environment. In 2019, Michael Tanana and colleagues (Tanana, Soma, Srikumar, Atkins, & Imel, 2019) tested the first ever application of a machine learning-based software that simulates a human patient. This system allows therapists and trainees to interact with a *bot* that provides appropriate answers based on the therapist's questions and reflections. This would not have been possible without the recent developments in context-aware recursive neural-networks (Vinyals & Le, 2015), which allow the *bot* to use coherent and plausible responses based on the therapist input. While it was clear to the participants of the study that the *bot* is not a realistic replacement for a human patient (Tanana, Soma, Srikumar, Atkins, & Imel, 2019, p. 10), systems such as this one could foreseeably be integrated into therapists' training, allowing them to improve their listening skills, retrieve concrete metrics from sessions in a reliable and fast manner, compare performance results over time and to others, and share the output with their supervisors who can use this information to critically develop strategies that will help them get better and grow more confident to deal with real patients.

The benefits of using text mining extend from the analysis of clinical session data to the analysis of non-clinical data. There is growing research on the use of topic modeling, a classification technique that looks for semantic similarity across groups of words. It has been used to identify topics covered during therapy sessions (Imel, Steyvers, & Atkins, 2015; Gaut, Steyvers, Imel, Atkins, & Smyth, 2017), predict therapeutic alliance ratings from the analysis of patient-therapist interactions in therapy sessions (Martinez, et al., 2019), predict psychosis from the analyses of Reddit forum posts (Rezaii, Walker, & Wolff, 2019), and identify personality and emotion from the analysis of Twitter posts (Quercia, Kosinski, Stillwell, & Crowcroft, 2011).[2]

Text mining can be an extremely useful tool during studies' design, research, and sample collection stages. Sara Santilli and Laura Nota (2017) wanting to retrieve the populational definition of the concept "courage", collected definitions of the concept provided by 1199 participants. They then analyzed the contributions using Latent Semantic Analysis and synthesized the answers into its shared words and relevant labels. Topic modeling can also be used to

[2] For thorough lists of examples of types of datasets explored and methodologies used, see Iliev, Dehghani, & Sagi (2015) and Calvo, Milne, Hussain, & Christensen (2017).

efficiently summarize themes in psychological literature, allowing researchers to aggregate large collections of text (articles, books, etc.) from different sources and apply the algorithm to classify them by genre and topic (Wang, et al., 2016). In addition, text mining can help researchers find appropriate samples in an efficient manner. Qiwei He (2018) was interested in developing a screening test for post-traumatic stress disorder (PTSD) by using lexical features in self-narratives. To avoid the time-consuming task of screening hundreds of possible participants in face-to-face interviews, she had volunteers write down their traumatic experiences and symptoms. These narratives were then analyzed, using sentiment analysis and topic modeling, to make classification decisions. In turn, those who were identified as displaying PTSD elements in the narratives were invited to an in-person diagnostic test and selected to participate in the study.

Application of Sentiment Analysis to In-Person Psychotherapy Session Data

Sentiment analysis can be used to identify particular subjects or track sentiment in textual data. It offers a convenient way to analyze large quantities of text and provide a summary of categories of interest. For instance, in the realm of psychiatric interest, it can be used as a pharmacovigilance tool by measuring the adverse effects of medications: we can collect references to medication names on social media and identify the sentiment associated to them (Korkontzelos et al., 2016, in Davcheva, 2018). In clinical psychology, the tracking of sentiment can be used to evaluate the success of a particular session. Althoff and colleagues (Althoff, Clark, & Leskovec, 2016) had patients rate individual sessions at the end of sessions via surveys. They noticed that patients who evaluated sessions as successful used more positive words and experienced a higher number of perspective changes throughout intra-sessions in comparison to patients who rated sessions negatively. Besides identifying subjects and tracking sentiment in text, the fact that the technique allows to analyze and summarize large quantities of data from multiple sources, sentiment analysis is sometimes able to uncover hard-to-find patterns. Elena Davcheva (2018) mined mental health forums to identify sentiment regarding mental health treaments, having found very interesting patterns related to various concepts such as family, medication, therapy, pets, sports, and medication for different types of mental disorders.

The analysis and validation of the application of sentiment analysis to psychotherapy corpora is still scant. One the one hand, published research that utilizes session data seems to focus on the application of topic modeling and other more sophisticated machine-learning prediction

models (Imel, Steyvers, & Atkins, 2015; Gaut, Steyvers, Imel, Atkins, & Smyth, 2017; Mieskes & Stiegelmayr, 2018; Martinez, et al., 2019). Also, if a sentiment analysis technique is used, it is often in its text classification approach, not in its lexicon-based variety. On the other hand, the research that integrates or focuses on sentiment analysis is not directly applied to in-person or face-to-face psychotherapy sessions, but rather to the analysis of mental health online forums (Davcheva, 2018), social media data from Twitter (Quercia, Kosinski, Stillwell, & Crowcroft, 2011) and Facebook, internet-delivered therapy (Provoost, Ruwaard, van Breda, Riper, & Bosse, 2019), and cell-phone text-based therapy (Althoff, Clark, & Leskovec, 2016). In addition, the lexicon-based approaches found in the literature can be grouped into two camps: they are performed using paid software (like LIWC), making access conditional; and those that are developed by the researchers themselves seem to be somewhat obscure, as access to neither the dictionaries nor the code used is made public in their papers for others to use or replicate. This opens the door to an exploration of the application of sentiment analysis to in-person psychotherapy data, with an exclusive focus on the performance of its lexicon-based approach. We used open-source software, along with publicly available lexicons provided in a text analysis package. Both can be accessed and used by anyone for free.

Even though one of advantages of using a lexicon-based approach to perform sentiment analysis is that the lexicons can be created and adapted by the researchers, tailoring them to address any particular theoretical construct of interest (Iliev, Dehghani, & Sagi, 2015, p. 6), for the scope of our research we aim to assess the performance of general purpose lexicons in order to evaluate how an out-of-the-box and theory-agnostic solution fares in comparison to human raters and how reliable they are to analyze psychotherapy session data.

3. Method

Design

Our exploratory study consisted of the evaluation of overall emotion and sentiment present in a patient's talk turns, from one complete in-person psychotherapy session, using a lexicon-based sentiment analysis approach. The evaluation was conducted 1) by human raters and 2) automatically by an algorithm using four different general purpose lexicons. Human ratings were evaluated both in terms of reliability, using inter-rater reliability and intra-class correlation measures. The four lexicon ratings were then compared to the human ratings to evaluate the performance of the sentiment analysis.

Materials

Session Data

For the purpose of this study we started out with eight English-speaking psychotherapy video sessions from a collection of training videos aimed at psychotherapists, psychotherapy trainees, and clinical psychology students. As we benefited from the help of a group of four transcription assistants (three psychology undergraduate students and one psychology graduate student), we devised a short set of guidelines based on Erhard Mergenthaler and Charles Stinson's (1992) transcription standards for psychotherapy sessions. These guidelines were set in a place not only to allow uniformity in the transcriptions but also to ensure the data was organized in a structure that would make computerized text analysis possible. For that reason, the data was structured in six columns: the first identified the talk turn number, the second identified the speaker (therapist or patient), the third included the talk turn text, the fourth and fifth columns identified the talk turns starting and ending times in a *hh:mm:ss.mm* format, and a sixth and seventh metadata columns were extracted post-transcription with notes and comments. The author evaluated and oversaw all transcription stages, ensuring fidelity and homogeneity across the corpora.

The eight sessions were transcribed verbatim, as faithfully to the spoken words of participants in the videos as possible. In instances where words were not clear for whatever reasons we classified those utterances in the text column inside square brackets as "[inaudible]". Moments of prolonged silences, displays of emotion, brief interpretations of unfinished words, and

vocabulary or grammatical peculiarities were identified in two separate metadata columns, designated “Notes” and “Comments”.

```
## Observations: 451
## Variables: 7
## $ doc_id   <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, …
## $ speaker  <chr> "Therapist", "Patient", "Therapist", "Patient", "Therap…
## $ text     <chr> "Okay, so hello.", "Hello.", "May begin... Hmm. What I …
## $ start    <time> 00:00:00.00, 00:00:02.62, 00:00:03.28, 00:00:22.67, 00…
## $ end      <time> 00:00:03.12, 00:00:03.16, 00:00:22.23, 00:00:23.48, 00…
## $ notes    <chr> "character(0)", "character(0)", "character(0)", "charac…
## $ comments <chr> "character(0)", "character(0)", "character(0)", "charac…
```

Figure 1: Structured session data

We identified each participant’s talk turn based on when a person started and stopped speaking. In instances where there was an overlap in speech, we broke up the speech by using ellipses, transcribing what the other person said in a new and separate text row, and resumed the other person’s talk turn in a new row by starting the text with ellipses.

```
##   doc_id speaker  text                                      start     end
##   <dbl>  <chr>    <chr>                                     <time>    <time>
## 1   165 Therapi… Well,  there's something about her... 15'39.97" 15'41.97"
## 2   166 Patient  Hmm-hmm.                                 15'42.15" 15'42.64"
## 3   167 Therapi… ... that's keeping you in a situatio… 15'43.51" 15'48.52"
```

Figure 2: Text sample with overlap in speech

We also annotated the starting and ending times of each turn, carefully documenting every audible utterance from participants, including hesitations, stuttering, false starts, nonverbal and incomplete words. Total session time was identified by computing the beginning of the first talk turn and the end of the last talk turn. For ethical and privacy reasons, participants names, referenced people and locations were removed and replaced by placeholders inside square brackets, for instance “New York” or “London” would become “[city]”, the name of a parent, a sibling, or a child would be replaced by “[father]”, “[sister]”, or “[daughter]”.

Since only one of the sessions was going to be evaluated by human raters, we randomly selected one for analysis using a random sampling function in software. The sampled therapy session was led by a female therapist, who applied Accelerated Experiential Dynamic Therapy (AEDT) to a female patient. AEDT is a therapeutic approach that focuses on an emotion-based transformational experience (Fosha, 2009). In the session the patient brought up themes related to her struggling and abusive marriage, her parents’ divorce and its lingering effects on the patient,

the patient's wrestling relationship with her mother, and a generalized concern about her daughter being exposed to a toxic domestic environment.

Questionnaires

After the process of transcribing, organizing the data in a computerized text analysis ready-format, data cleaning, and ensuring it was converted to a universal text encoding format (UTF-8) to avoid compatibility issues, we created an interactive web-based application that was used as a questionnaire form for our participants to access and fill out. We based our application programming code on Markus Steiner and colleagues' (Steiner, Phillips, & Trutmann, 2019) ShinyPsych Survey application, which we adapted in order to fit our particular needs.3 For ease of use and to maximize respondent rate, we enabled our application to run on personal computers and on mobile devices, such as smart phones and tablets. After each rater finished filling out the questionnaire, a text file was sent to the researcher with the answers, along with some demographic data and respondent ID for verification.

The total number of patient talk turns were 224. However, since some of them were duplicates (instances where the patient said exactly the same thing in different occasions), we trimmed the dataset into 176 unique talk turns and 4 representing all the other repeated instances. Those four represented talk turns where the patient said "Okay.", "Hmm.", "Hmm-hmm.", and "Yeah.". Due to the number of talk turns subject to evaluation, we split the dataset into two questionnaires, each with 90 talk turns for evaluation on two dimensions – sentiment and emotion. We randomized the order of the talk turns, so that they would not appear in sequential order to raters.

Once participants accessed the questionnaire they were shown information about the nature and objective of the study. After they gave their consent to participate, they were shown a tutorial with a set of instructions on how to perform the rating tasks, by practicing the manipulation of the tools on the application. As participants proceeded and started to take the questionnaire, they were shown a text representing the talk turn. Text could be as long as one word or a set of sentences, depending on the length of the talk turn being evaluated.

[3] To access a sample copy of the questionnaires used in this study please visit: https://goncalomveiga.shinyapps.io/Questionnaire1/.

The first task comprised the rating of sentiment. Participants were prompted to read a text ("Please consider the following text:") and underneath it they would have a slider, which they could manipulate from any value between -5 to 5. Raters were informed that if the text was neutral to just make sure the lever was on the 0 position and press Continue. If the text was negative participants were instructed to adjust the slider to the left of 0, to whichever value they thought represented the intensity. Finally, if the text was positive they were instructed to move the lever on the slide to the right of 0, to whichever value they thought represented the intensity.

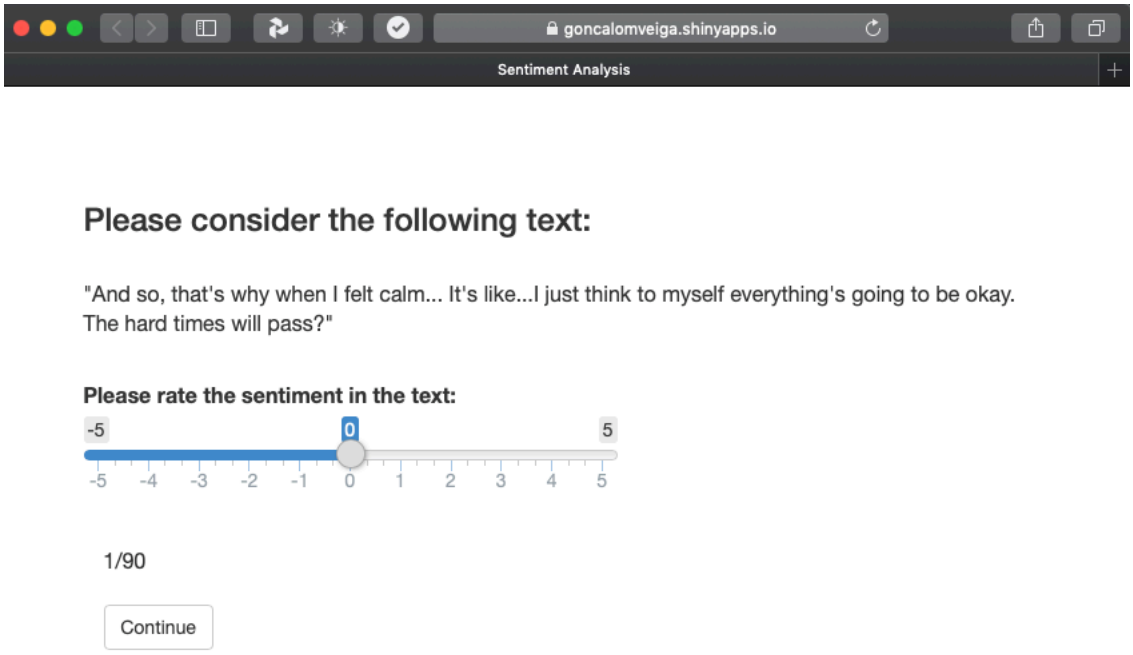

Figure 3: Rating Sentiment in text

The second task involved identifying emotion in text from a list of possible emotions. The emotions from the list are extracted from the NRC Lexicon, which is based on the American psychologist Robert Plutchik's Wheel of Emotions4. The NRC Lexicon was created and is maintained by a group of experts of the National Research Council of Canada (Mohammad, 2006). The original NRC list includes a set of eight emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) and a set of two sentiment ratings (positive and negative).

[4] Plutchik conceptualized emotions in a mandala-like structure, where diametrically opposed emotions are placed on opposite sides of the wheel, and similar emotions emanate from the center of the diagram to the outer edges toward each of the eight basic emotions he identified. To visualize the model consult (Plutchik, 2001, p. 349).

For the purpose of our study, we filtered out the sentiments ratings from the NRC Lexicon, as we will be using three other lexicons to evaluate sentiment, and we added "None of the above" in case no emotion was identified by the participant. If more than one emotion was identified in the text, participants were prompted to select the prevalent one.

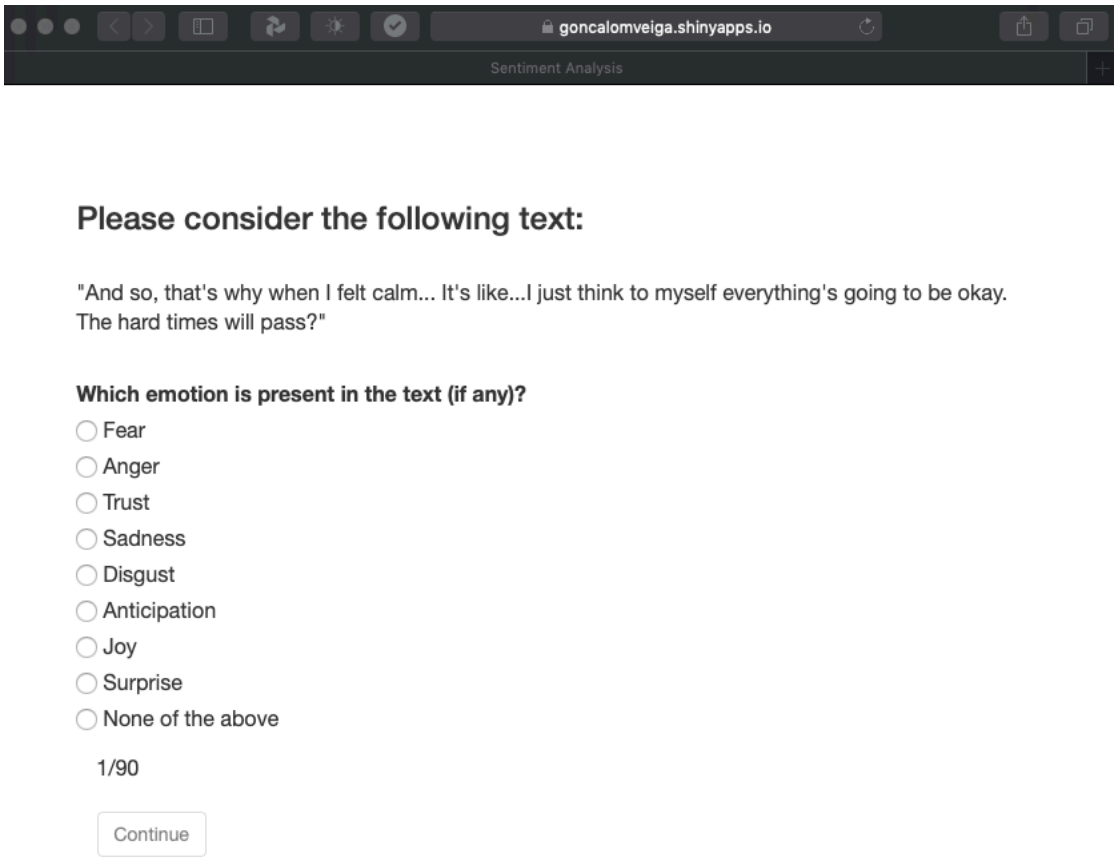

Figure 4: Identifying emotion in text

Software Applications and Packages

We used Microsoft Excel to transfer the transcription from a Word text file to a spreadsheet format in order to structure the data into rows and columns. The majority of the data cleaning, processing, analysis, and visualization steps was performed using the open-source statistical programming language and environment R (R Core Team, 2018). The programming language comes prebuilt with base packages but in order to perform our analysis we augmented it with other authors' packages. We predominantly used the tidyverse collection of packages developed by Hadley Wickham (2017) to perform most of the data cleaning, wrangling, exploration, and visualization. To perform the sentiment analyses we used Julia Silge and David Robinson's (2016)

tidytext package along with their manual Text Mining with R (2017). To calculate inter-rater reliability and intra-class correlations we used the rel (Martire, 2017), the psych (Revelle, 2018), and irr (Gamer, Lemon, & Singh, 2019) packages.

Lexicons

The general-purpose lexicons used in this study were made available through the tidytext package. The first lexicon used was NRC, which includes ratings for 13901 terms and rates terms on sentiment (positive and negative) and on eight emotions, some with overlap.

```
##    NRC          `Count of terms`
##    <chr>                   <int>
##  1 anger                    1247
##  2 anticipation              839
##  3 disgust                  1058
##  4 fear                     1476
##  5 joy                       689
##  6 negative                 3324
##  7 positive                 2312
##  8 sadness                  1191
##  9 surprise                  534
## 10 trust                    1231
```

Figure 5: NRC lexicon count of terms

The second lexicon used was Bing, developed by Bing Liu and colleagues (Bing, 2019). It contains 6786 terms, 2005 are rated positive and 4781 negative.

```
##   Bing     `Count of terms`
##   <chr>               <int>
## 1 negative             4781
## 2 positive             2005
```

Figure 6: Bing lexicon count of terms

The third lexicon used was Loughran-McDonald, developed by Tim Loughran and Bill McDonald (Loughran & McDonald, 2011). The lexicon was constructed for use in finance and law. It contains 4150 terms and identifies four emotions and sentiment (positive and negative).

```
sentiment    `n()`
##   <chr>        <int>
## 1 constraining   184
## 2 litigious      904
## 3 negative      2355
## 4 positive       354
## 5 superfluous     56
## 6 uncertainty    297
```

Figure 7: Loughran-McDonald Lexicon count of terms

The fourth lexicon used was AFINN, developed by Finn Arup Nielsen (Nielsen, 2011). It includes 2477 terms and terms are rated between -5 to 5.

```
##    AFINN `Count of terms`
##    <dbl>            <int>
##  1    -5               16
##  2    -4               43
##  3    -3              264
##  4    -2              966
##  5    -1              309
##  6     0                1
##  7     1              208
##  8     2              448
##  9     3              172
## 10     4               45
## 11     5                5
```

Figure 8: AFINN lexicon count of terms

Participants

The questionnaires were made available to the public for a period of 30 days, during the months of September and October of 2019. A total of N = 52 raters were recruited by email, on social media, and through professional and academic networks both in the USA and abroad. The raters identified themselves as N = 28 female, and N = 24 male. 36 are native English speakers, while 16 use English as a second-language. The average age was 40.96 years, with a standard deviation of 14.06 years.

Raters were assured their participation was voluntary, that their personal information would be anonymized, and that filling out the questionnaire did not pose any health risks. Participants were not promised nor received compensation of any type for their participation in the study.

```
##    Total raters  Mean Age Mode Age Median Min Max Range  Std Deviation
## 1            52     40.96       36     37  18  74    56       14.06283

##   gender language                   Total `Mean Age`
##   <fct>  <fct>                      <int>      <dbl>
## 1 male   native speaker                15       38.7
## 2 male   English as second-language     9       37.6
## 3 female native speaker                21       44.8
## 4 female English as second-language     7       38.7
```

Figure 9: Raters demographics

Procedure

Computerized Sentiment Analyses

All four lexicons used in this study are based on unigrams, that is, single words. For that reason we had to use a bag-of-words method, which consisted in transforming each talk turn into a collection of single words or tokens. We tokenized our text, while still retaining information on the talk turn each word belongs to.

Subsequently, we merged each of the lexicons datasets with a separate copy of the patient talk turns dataset, creating a distinct dataset for the NRC, Bing, Loughran-McDonald, and AFINN lexicons. This merge was performed based on the words that the patient dataset has in common with each of the lexicons. Since each of the lexicons does not contain all the words in the English vocabulary, we performed a merge of the intersection of the terms, not an union. After the three new datasets were created, we performed the sentiment analysis on each one of them. For three of the lexicons we applied additional differentiated steps, which we will mention below.

NRC Sentiment Analysis

Due to the fact that we were exclusively interested in identifying emotions in the talk turns using the NRC lexicon, we filtered out the "Positive" and "Negative" sentiment ratings from the lexicon before merging the two datasets, leaving only the eight emotion ratings for analysis.

The NRC lexicon has the ability to identify more than one emotion both per term and per talk turn. To illustrate the former, a term such as "abandoned" is identified as possessing the following emotions: anger, fear, sadness. However, we were interested in knowing if it can reliably identify an emotion that is also identified by human raters. Since the human ratings for emotion is based on the prevalent emotion in the talk turn, we calculated the human mode to represent the emotion per talk turn. As we cannot compute the inter-rater reliability measures between a single value and a vector of values, we executed the following decision: 1) If the human rating mode for a particular talk turn is X and there is only one emotion detected by the NRC lexicon for the same talk turn, we did not make any changes to the NRC rating for the talk turn; 2) If the human rating mode for a particular talk turn is X and the emotion detected by the NRC lexicon for the talk turn is a vector such as X, Y, Z, we decided that the NRC detected the same emotion as the human

ratings mode, keep X as the NRC talk turn rating and remove Y and Z; 3) If the human rating mode for a particular talk turn is X and the emotion detected by the lexicon for that same talk turn is a vector such as Y and Z, we decided that there was no match between the human ratings mode and the NRC lexicon and we proceeded to randomly select one of the ratings to represent the NRC emotion and delete the remaining.

Bing Sentiment Analysis

We applied the general steps for the Bing sentiment analysis but removed two stopwords ("Right", "Yeah"). Stopwords are terms that are present in language but whose interpretational value is questionable and has the potential to skew the results erroneously. Unlike all the other lexicons, Bing has a rating for those two terms. As we did not perform a general stopwords removal (filtering them out is a standard step) to keep the data for analysis as integral as possible, these two terms were removed as they were prevalent across the session. If multiple ratings per turn occurred, we computed the mode to retrieve the most prevalent value.

Loughran-McDonald Sentiment Analysis

The Loughran lexicon identifies the sentiment (positive and negative) and four emotions in terms. As we were only interested in its value as a sentiment dataset, we filtered out the emotions from the analysis. If multiple ratings per talk turn were found, we computed the mode for those values.

AFINN

AFINN's sentiment analysis output per talk turn can be a single value or a vector of values, depending on how many terms it found and rated. As the data is ordinal, if more than one rating was found we computed the median to represent the average result per talk turn. We also removed two stopwords ("Yeah", "Like") for our analysis.

Human Raters Sentiment Analysis

We collected and compiled the raters answers and separated them into two groups. One group contained all the ratings for emotion detection and the other group included all the sentiment ratings. In order to compute the inter-rater reliability between human raters and the inter-rater reliability between human raters and lexicons, we needed to 1) convert the data to the appropriate variable scale and 2) compute the appropriate average for each measurement scale.

The human ratings for comparison to the NRC sentiment analysis were already in a nominal level of measurement, so we just calculated the human rating average per talk turn by computing the mode.

For the analysis of sentiment we took a two-pronged approach. For a comparison to the AFINN lexicon, we converted the human ratings from a continuous scale to an ordinal level of measurement one by converting the continuous values to a - 5 to 5 ordinal scale and computed the human rating average per talk turn by calculating the median. For the comparison to the Loughran-McDonald and Bing lexicons, we converted raters responses from the continuous scale to a nominal one: rating values < 0 were converted to "negative", ratings of 0 were converted to "neutral", and values > 0 were converted to "positive". Finally, for the inter-rater reliability calculation for those two nominal-based lexicons we calculated the mode to represent the human rating average of sentiment per talk turn.

Inter-rater Reliability (IRR)

The assessment of inter-rater reliability (IRR) allows us to quantify the degree of agreement between two or more raters making independent ratings about subjects (Hallgren, 2012). In our study, the subjects were the talk turns that were evaluated on emotion and sentiment both by human raters and different lexicons. Due to the variety of our data in terms of scales of measurement (nominal and ordinal), different types of rater groups (human-human, human-lexicons), and whether each rater evaluated each talk turn or a subset of talk turns (missing values), we had to use several statistical formulas to analyze the agreement between raters.

Krippendorff's Alpha

We used Krippendorff's alpha throughout all the IRR calculations as it can be applied to all scales of measurement (nominal to ratio), it is able to deal with missing values, and it can be used to evaluate the reliability for two or more raters (Hayes & Krippendorff, 2007). The alpha value range from – 1 to 1, where 1 finds perfect agreement between raters and 0 finds no agreement. According to Klaus Krippendorff (2004, pp. 241 - 242), we should rely only on variables whose ratings achieved an alpha > 0.8 to make generalizations, and alpha values between 0.667 and 0.8 to draw tentative conclusions. The same is applied to the analysis of the confidence intervals, which are extracted via bootstrap sampling.

Cohen's Kappa

Cohen's kappa is used to assess the IRR for nominal variables. It can only be applied to instances where we are measuring the agreement between two raters and does not allow for missing values in the data. The kappa values range between - 1 and 1. If a kappa value is negative there is disagreement, and if the value ranges between 0 to 0.2 it indicates slight agreement, 0.21 to 0.4 fair agreement, 0.41 to 0.6 moderate agreement, 0.61 to 0.8 substantial agreement, and 0.81 to 1 indicates strong or perfect agreement (Hallgren, 2012).

Fleiss' Kappa

Fleiss' kappa is an extension of Cohen's kappa but it allows to evaluate agreement for three or more raters. However, like Cohen's kappa it does not allow for missing values and it can only be used on nominal variables. The Interpretation of Fleiss' kappa values is the same as Cohen's kappa.

Inter-class Correlations (ICC)

Inter-class correlations (ICC) allow us to assess inter-rater agreement from ordinal to ratio levels of measurement. Unlike Cohen's and Fleiss's kappas, which calculate IRR by quantifying disagreement in an all-or-nothing approach, ICCs calculate IRR by incorporating the magnitude of disagreement in the calculation (Hallgren, 2012). In this way, larger magnitude disagreements

in ratings generate lower ICC scores than in scenarios where the magnitude is lower. ICC values less than 0.4 indicate poor agreement, 0.40 to 0.59 fair agreement, 0.60 and 0.74 fair agreement, and values between 0.75 to 1 indicate excellent agreement.

Following Shrout and Fleiss' (Shrout & Fleiss, 1979) terminology, we will assess human-human ICC via ICC(2,1): two-way/random raters, measuring agreement, and single unit. To assess the human-lexicons we will use ICC(3,1): two-way/fixed raters, measuring consistency, and single unit.

4. Results

In this section we present the different sentiment analyses of the session data performed by the group of human raters and by the four lexicons. We start by providing some contextual data on the session. Then we introduce the results from the human and NRC sentiment analyses for detecting emotion and, subsequently, the results from the human and remaining lexicon sentiment analyses.

Session Data

In total there were 451 talk turns in the session. We registered 224 patient's talk turns and 227 from the therapist. The total session time was 52 and 15 seconds. The patient had an average talk turn duration of 7.73 seconds, having spoken a total time of 28 minutes and 53 seconds in the session. On average the patient spoke 81.29 characters per turn and 16.11 words per turn. The patient's longest turn word-wise was 336, and the shortest was one word long. On the therapist side, the average talk turn duration was 3.86 seconds, for a total spoken time of 14 minutes and 28 seconds across the session. The therapist spoke on average 39.37 characters per turn, 7.86 words per turn. The longest word-wise talk turn for the therapist was 54 words and the shortest one word long. In total 5401 words were spoken in the session.

Rating Emotion

Human Raters Emotion Ratings

The total human ratings of emotions was 8294, including None of the above responses. Even though not all of the 52 raters rated emotion across the session, the response rate for emotion was still very high, at 88.61%. The maximum number of evaluations per rater was 180, the minimum was 87, and the average for all participants was 159.5 ratings on emotion. Each talk turn was evaluated on emotion on average by 46.08 raters, with a maximum number of ratings of 49 and a minimum of 42. Of the 180 talk turns, 178 had a single mode and two had a bimodal rating.

In the two talk turns that received a bimodal score (Figure 10), the first one had 12 ratings both for “disgust” and “sadness”, while the second one received 17 ratings for “anger” and “sadness”.

```
## [1] "Hmm. She was just very uninvolved. She was never interested in learni
ng anything beyond what she already knew. You know..."

## [2] "And so, that makes me sad and mad. And... But then I just think of wh
at our relationship is today and how it's me pursuing her, me trying to keep
that relationship together. Because..."
```

Figure 10: Two talk turns with bimodal ratings on emotion

After assessing the mode for each talk turn, our results followed the same general pattern of the total ratings, with a majority of ratings being None of the above (95), Sadness (41), followed by Fear (13), Joy (8), Anger (7), Anticipation (7), Surprise (4), Trust (3), and Disgust (2). Figure 11 shows a representation of the mode ratings across the session, allowing us to observe how participants rated the talk turns per emotion across in time. We filtered out None of the above ratings as they do not identify any emotion and are very numerous.

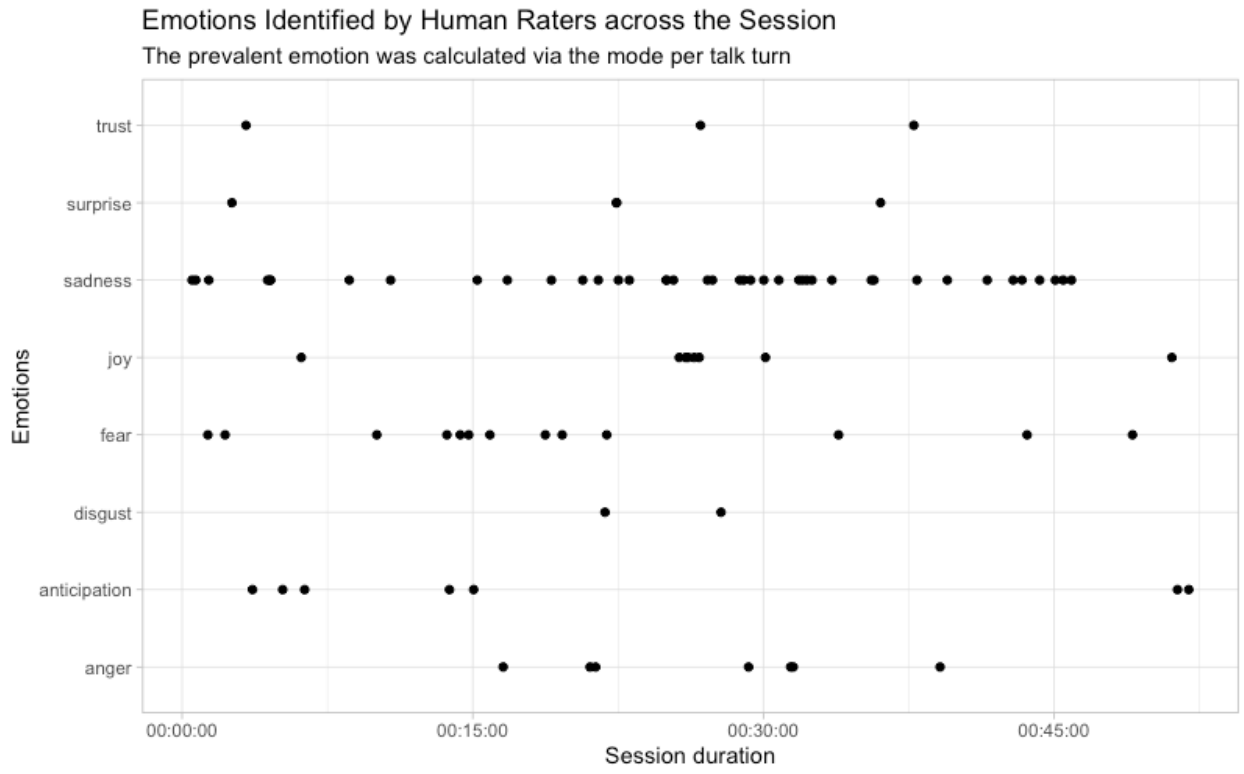

Figure 11: Emotions identified by human raters across the session

As we can observe from Figure 11, Sadness has been identified extensively across the session, except right near the end. Joy was selected close to the beginning of the session, at mid-point, and near the end. Fear appears to be more concentrated in the first half of the session. Anger is more pervasive around the middle of the session, without any observations either in the beginning or at the end. Just like Anger, Disgust is also only identified at the middle. On the other hand, Anticipation seems to be distributed around the extremes of the session, with a larger cluster in the beginning and two observations at the end.

NRC Lexicon Emotion Ratings

After applying the NRC sentiment analysis to retrieve emotions, the lexicon sorts out the terms it does not share in common with the session text data. By plotting the ten most frequently occurring words in the text (Figure 12), we can see that Feeling, Grow, and Marriage are the most prevalent across the session with at least 30 instances across the session. Verb-wise we can observe Feeling, Grow, Share, Deal, and Struggle. Noun-wise, we have Marriage, Divorce, and Daughter, which might suggest these were the most frequently discussed topics throughout the session.

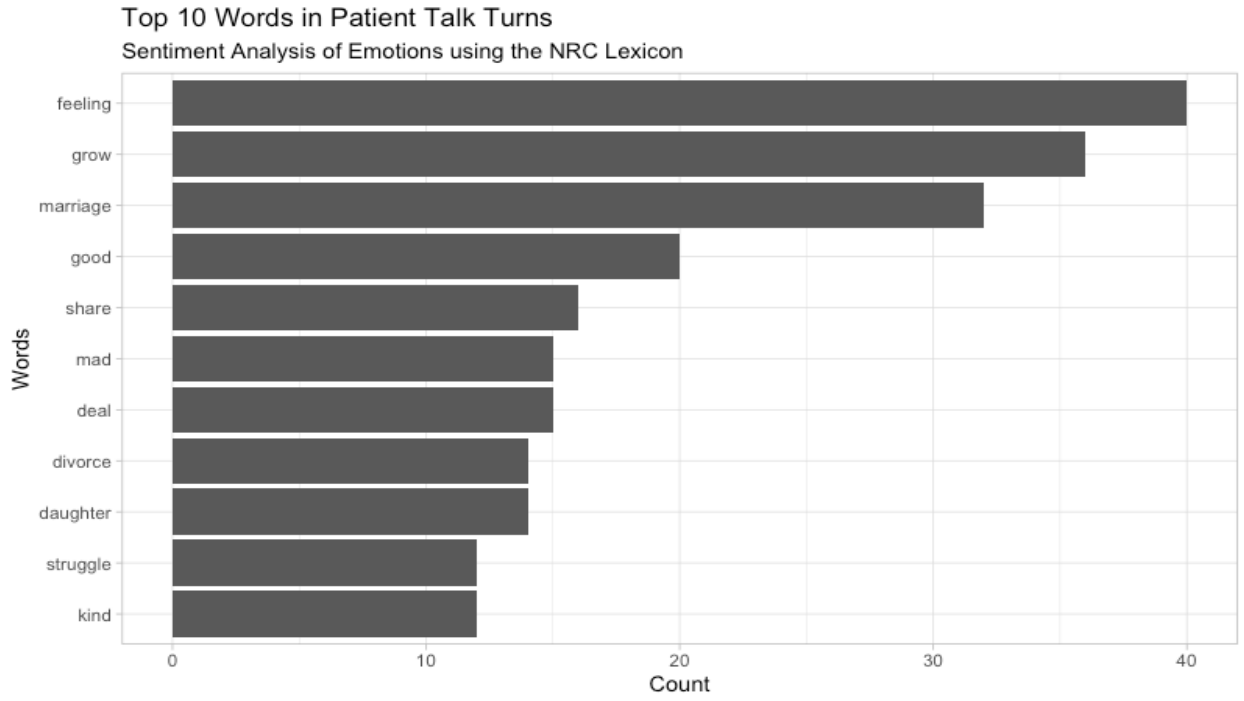

Figure 12: NRC Top 10 words in patient talk turns

The NRC ratings are more numerous than the human's, as the lexicon can identify more than one emotion in each term and per talk turn. The most numerous emotion identified was Trust (68), closely followed by Joy (64) and Anticipation (63). Fear was found 36 times, Sadness 28, and Anger 27. Disgust and Surprise have the smaller frequency count with 22 and 20 ratings, respectively.

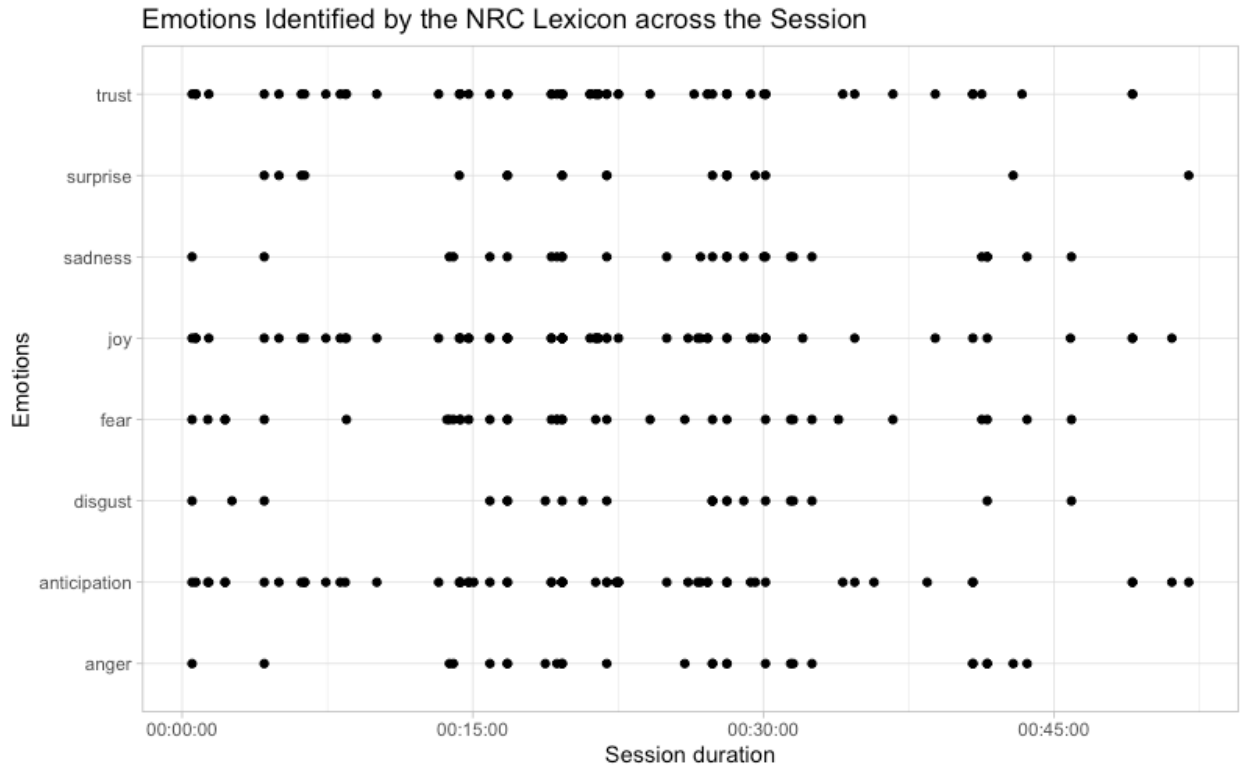

Figure 13: Emotions identified by the NRC lexicon across the session

In Figure 13, we plotted the emotions identified by the NRC lexicon across the session. Unlike in the human ratings' emotion plot, these are not mode values which make them more numerous and overlap across the timeline. Disgust, which had not been identified as a prevalent emotion in any of the talk turns by humans, has been identified throughout most of the session. Interestingly, the two instances where human raters identified Disgust as the prevalent emotion seem to match the edge of two dense clusters of the NRC's ratings for disgust. Anger has been identified in many instances across the session but, again, with a denser frequency around the middle, which matches the human mode. Trust, Surprise, and Joy have much higher frequencies than the human ratings. On the other hand, Sadness seems to match visually in range to NRC's ratings, even though the NRC's ratings are more spotted in certain portions of that range.

Assessing Inter-rater Reliability for the Human and NRC analysis

For the human-human emotion ratings, Krippendorff's alpha for the 52 raters was α = 0.226, CI = 0.20 to 0.24, confidence level = 0.95, for 180 observations/talk turns. If we analyze the subset of human raters (n = 27 raters) with no missing values in their emotion ratings, Krippendorff's alpha was α = 0.227, CI = 0.20 to 0.25, confidence level = 0.95. Fleiss' kappa does not allow missing values in the calculation, so we measured Fleiss' kappa for n = 27 raters, which was κ = 0.229 for the 180 observations. Human-human inter-rater reliability was low according to Krippendorff's alpha and according to Fleiss' kappa we have a Fair measure of agreement.

To calculate the human-NRC inter-rater reliability we tried two different calculations, the first calculation included all instances where both raters and NRC did not find an emotion, while the second included only the eight emotions on both sets.

The first calculation involved leaving in instances where human raters did not identify any emotion in text (None of the above) and where the NRC lexicon did not find any term to rate as an emotion. For the latter, we searched for those NA rows and labelled them None of the above. This made it possible to calculate the IRR on 180 observations. Krippendorff's alpha for human-NRC for 180 talk turns was α = 0.296, CI = 0.18 to 0.37, confidence level = 0.95. Cohen's kappa for human-NRC for 180 talk turns was κ = 0.307, CI = 0.19 to 0.41, confidence level = 0.95. In other words, low or unreliable according to Krippendorff's alpha and Fair according to Cohen's kappa.

For the second calculation, all None of the above ratings for human and NRC ratings were removed, which trimmed the analysis to 47 talk turns. Calculating Krippendorff's for the human-NRC ratings for the 47 observations, we retrieved an alpha α = 0.444, CI = .26 to 0.59. And calculating Cohen's kappa for the human-NRC for the 47 observations retrieved a kappa of κ = 0.452, CI = 0.27 to 0.63, confidence level = 0.95. The trimmed dataset without the "None of the above" ratings performed slightly better, achieving a Moderate level of agreement according to Cohen's kappa but still retaining an unreliable value by Krippendorff's standards.

Rating Sentiment – Nominal Lexicons

Human Raters – Nominal Interpretation of Sentiment

The total human ratings of sentiment was 8370, with a response rate of 89%. On average, each participant rated 161 questions, with a minimum of 90 and a maximum of 180 questions. Each talk turn was rated on average by 46.5 raters, with a minimum of 44 raters per talk turn and a maximum of 49. The count of ratings was as follows: 1993 were positive, 3706 were negative, and 2671 were rated as neutral. All the talk turns had one mode. The human ratings mode values for the session were: 86 negative talk turns, 58 neutral, and 36 positive.

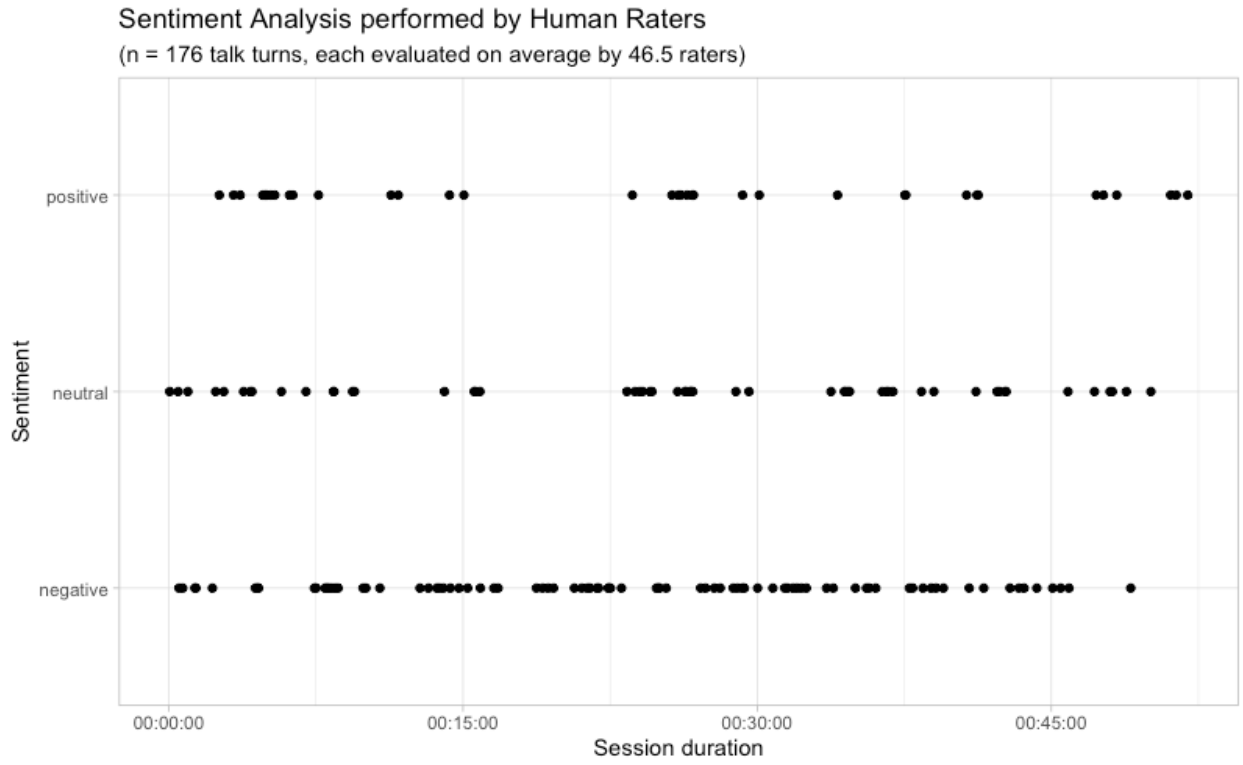

Figure 14: Sentiment Analysis performed by Human Raters

In Figure 14 we can visualize the distribution of sentiment ratings performed by human raters, according to the mode values. We can observe that the session starts off with neutral and negative ratings, but ends with positive ratings. In fact, with the exception of one negative rating at minute 49, the previous negative rating occurred at the end of minute 45. The longest consecutive period of negative ratings occurred between minutes 16 and 23, suggesting that was a period when the patient used numerous emotionally charged words. There is another period of

consecutive negative ratings, between minute 30 and 33, but which is shorter in duration relative to the one just previously mentioned. As raters identified more than double of negative talk turns in relation to positive ones, it is not surprising that positive talk turns are sporadic and far between in portions of the session. Neutral ratings seem to cluster in the beginning of the session and in the middle, having a relatively more wide spread concentration of values close to the end.

The Bing Lexicon

For the Bing Sentiment Analysis we had to perform a stopwords removal for a couple of words that are ambiguous and have the ability to skew the results of the analysis. As we can see in Figure 15, after we apply the Bing lexicon to our data, there is an overwhelming frequency of two words that matched the terms in the lexicon. While the term Like can have a positive rating, its use as a verbal filler disqualifies it for use in analysis. The same principle applies to the word Right, whose ambiguous use as either a positive term or a filler makes it unfit for use in the analysis.

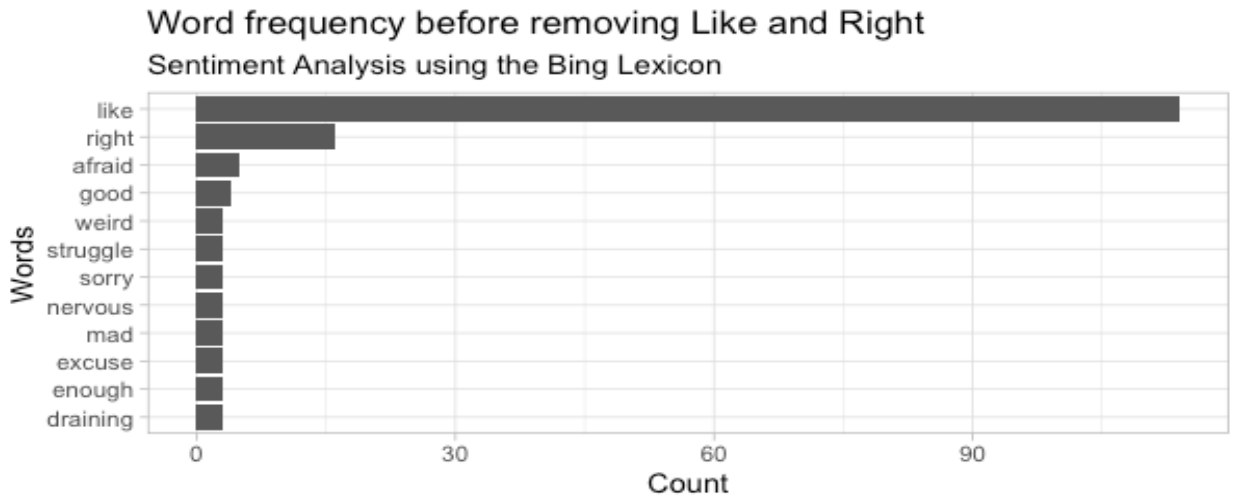

Figure 15: Word frequency before removing Like and Right

Bing's lexicon does not possess neutral terms, reducing the sentiment analysis to a positive and negative evaluation of the talk turns. In this evaluation, Bing identified 66 negative and 41 positive terms in 75 talk turns. It seems negative ratings have a wider range of values across the session only by a slight margin, revealing the session started and ended with negative ratings. It identifies two periods of time with a long prevalence of negative ratings, between minute 31 and 36, and minute 41 to 48. The longest sequence of positive ratings occurs for five consecutive occasion, for less than a minute, between minutes 20 to 21. In is also worth noting that there were

instances in the session for which the lexicon evaluated no talk terms, the longest one being between minute 10 and close to minute 14.

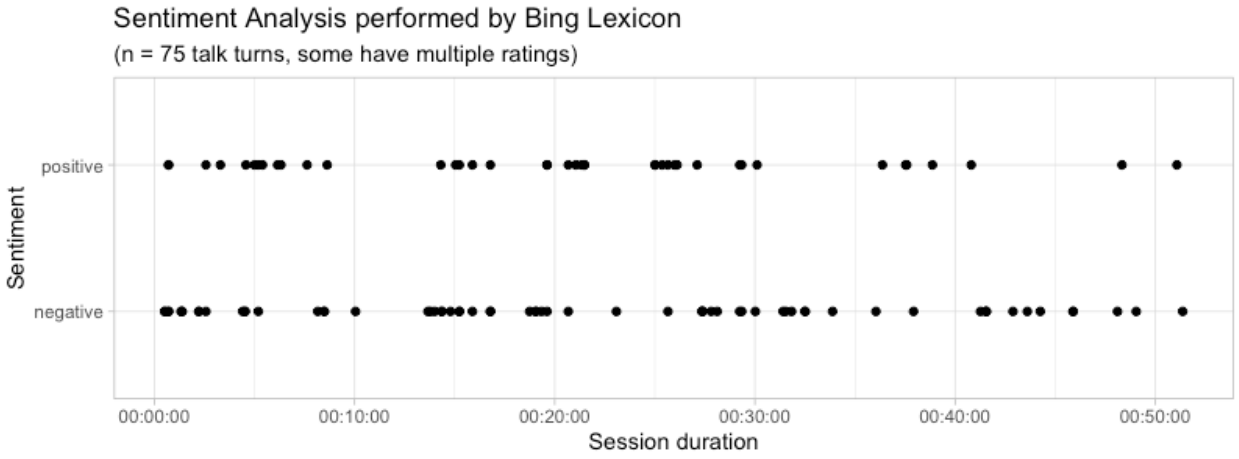

Figure 16: Sentiment Analysis perform by Bing Lexicon

The Loughran-McDonald Lexicon

The frequency count of the terms Loughran-McDonald lexicon found in the session data (Figure 17: Top 10 Words using the Loughran-McDonald Lexicon)indicate that the top terms matched with the patient's speech reflect a sense of hesitation. The fact that terms such as Could, Probably, Maybe, Doubt, and Almost appear as the most frequent seem to support that impression. It is also worth noting that the terms Divorce and Divorced matched as different entities, and they were both identified as two of the most frequently occurring terms in the text.

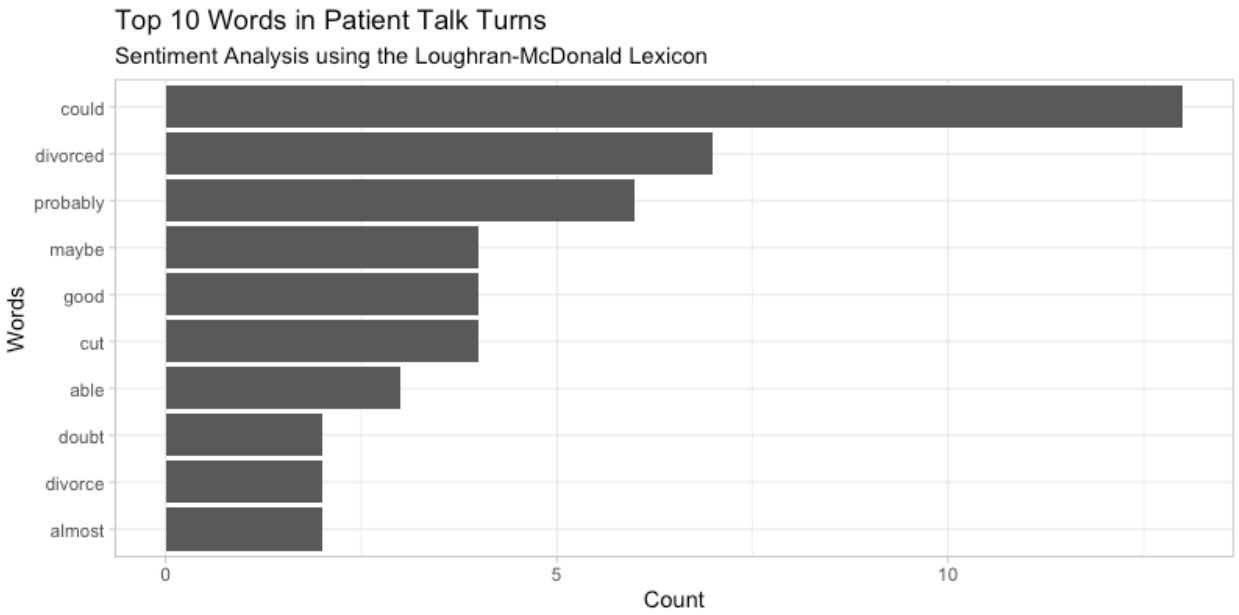

Figure 17: Top 10 Words using the Loughran-McDonald Lexicon

The Loughran-McDonald lexicon detected 35 negative and 13 positive words, across 39 turns. The low frequency of words matched with the talk turns text data make the ratings look very sparse across the session. In fact there are numerous portions of session time that have no rantings, the longest being between minute 1 and 5, and minute 10 to 14. The analysis of sentiment using the lexicon would suggest that the session started and ended negatively. Positive ratings seem to cluster more densely around minute 5, and then more sparsely around minutes 27 and 45. The longest consecutively number of negative ratings (13) occurred between minutes 14 and 23.

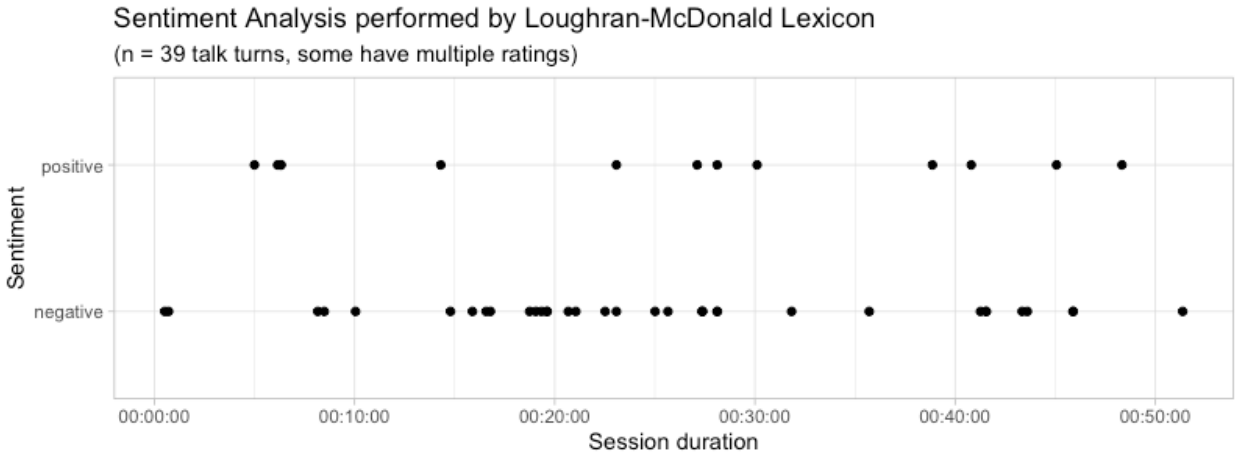

Figure 18: Sentiment Analysis performed by Loughran-McDonald Lexicon

Assessing Inter-Rater Reliability for the Human Raters and Nominal Lexicons

The inter-rater reliability for the sample of human raters was evaluated by Krippendorff's alpha on n = 52 and n = 41 raters on 180 observations/talk turns, and by Fleiss' kappa on n = 41 raters. Krippendorff's alpha for the total number of raters was $\alpha = 0.337$, with a CI = 0.32 to 0.37. confidence level = 0.95. For the subset of raters with no missing values in their ratings, Krippendorff's alpha performed very similarly $\alpha = 0.341$, CI = 0.30 to 0.38, confidence level = 0.95. Fleiss' kappa for the 41 raters was $\kappa = 0.34$. The results indicate that the group of raters has a low reliability score in Krippendorff's interpretation, and according to Fleiss's kappa the result suggest a Fair agreement between human raters.

For the assessment of agreement between human ratings and the Bing lexicon ratings, we used Krippendorff's alpha and Cohen's kappa to evaluate the level of agreement between the human ratings mode and Bing's scores. Krippendorff's alpha for 180 observations/talk turns with missing values was $\alpha = 0.479$, CI = 0.24 to 0.67 and for the subset of ratings with no missing values (75 observations) was $\alpha = 0.479$, CI = 0.22 to 0.61. Both used the default confidence level = 0.95, with 100 bootstrap sampling. We also assessed the agreement level between the human ratings mode and Bing's ratings by computing Cohen's kappa at a confidence level = 0.95. Cohen's kappa score was $\kappa = 0.48$, CI = 0.27 to 0.69. Krippendorff's alpha scores indicate that there is low reliability between human ratings and Bing ratings. On the other hand, according to Cohen's kappa the inter-rater reliability between the two groups of ratings denotes a moderate agreement.

To evaluate the IRR between the human ratings mode and Loughran-McDonald's ratings we also used Krippendorff's alpha and Cohen's kappa, both with a confidence level = 0.95. For the total number of observations n = 180 but with missing values, Krippendorff's alpha assessment was $\alpha = 0.519$, CI = 0.21 to 0.83. For the evaluation with the subset of data with no missing ratings, n = 39, Krippendorff's alpha value was $\alpha = 0.51$, CI = 0.18 to 0.77. Cohen's kappa, which only calculates the agreement between raters with no missing values, the kappa score was $\kappa = 0.513$, CI = 0.17 to 0.85. Even though both Krippendorff's alphas upper confidence intervals have values above CI = 0.77, the lower CIs have values far below CI = 0.667, which is for the cutoff point for appropriate reliability according to Krippendorff (Krippendorff, 2004). Cohen's kappa suggest moderate agreement between human ratings and Loughran-McDonald's lexicon ratings.

Rating Sentiment – Ordinal Lexicon

Human Raters – Ordinal Interpretation of Sentiment

The analysis of sentiment based on median values allow us to observe the polarity and intensity of the ratings per each talk turn. A visual inspection of Figure 19 allow us to observe the noticeable contrast between the number of negative ratings (75) and positive ratings (27). Some gaps along the zero axis somewhat conceal there is also a plethora of neutral ratings (74). The session seems to have started off negatively, remained negative throughout but ended on a sequence of three positive ratings. According to human raters, there were only two short periods of mostly positive moments across the session. The first one occurs between the end of minute 4 until half-way through minute 6. It is peculiar how that streak of positive ratings start off with a rating of 0, changed to 1, and ascended to 2 and ended on two consecutive ratings of 3. The second cluster of positive ratings occurs between minute 25 and 27. The ability to assess the session on intensity is an interesting contrast from negative and positive-only lexicons.

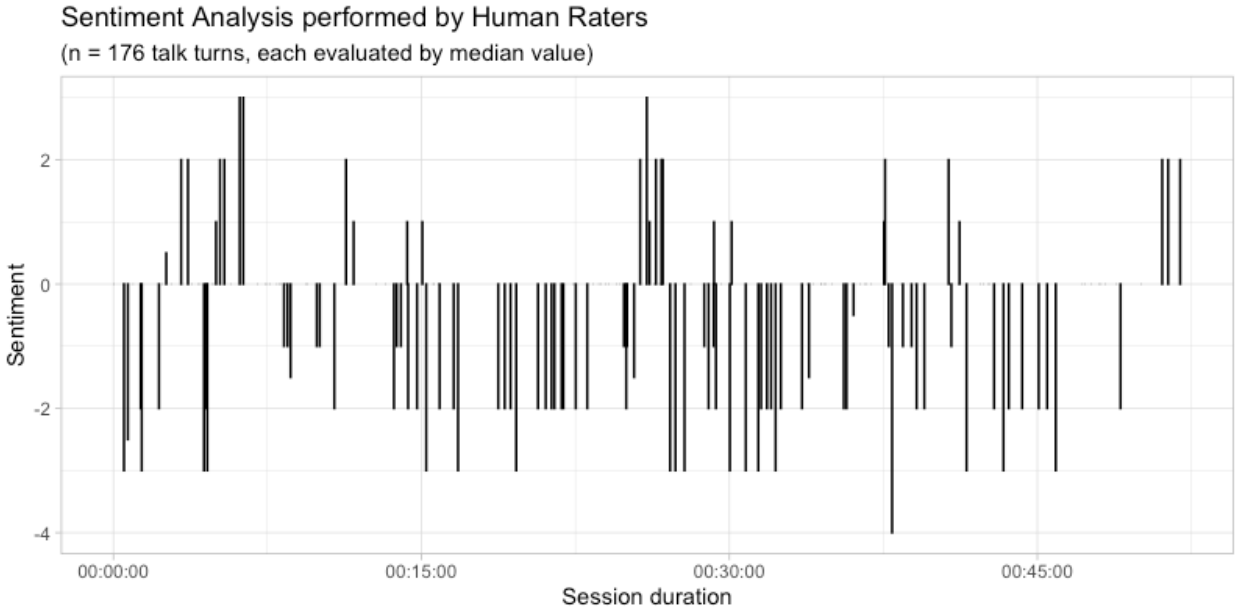

Figure 19: Sentiment Analysis performed by Human Raters based on median values

The AFINN Lexicon

The AFINN lexicon evaluated the session on 164 terms, across a total of 88 talk turns. As multiple talk turns had more than one rating we calculated the average value by computing the median. Most of the frequent words in Figure 20 seem to be negative. The two nouns, Afraid and Alone, along with the determiner No, and the verb Want seem to suggest that the lexicon identified

needs and negative emotions in the patient's speech. Even though there is a positive adjective in the list, Good, the verbs Cut, Feeling, and Want seem to offset toward negativity.

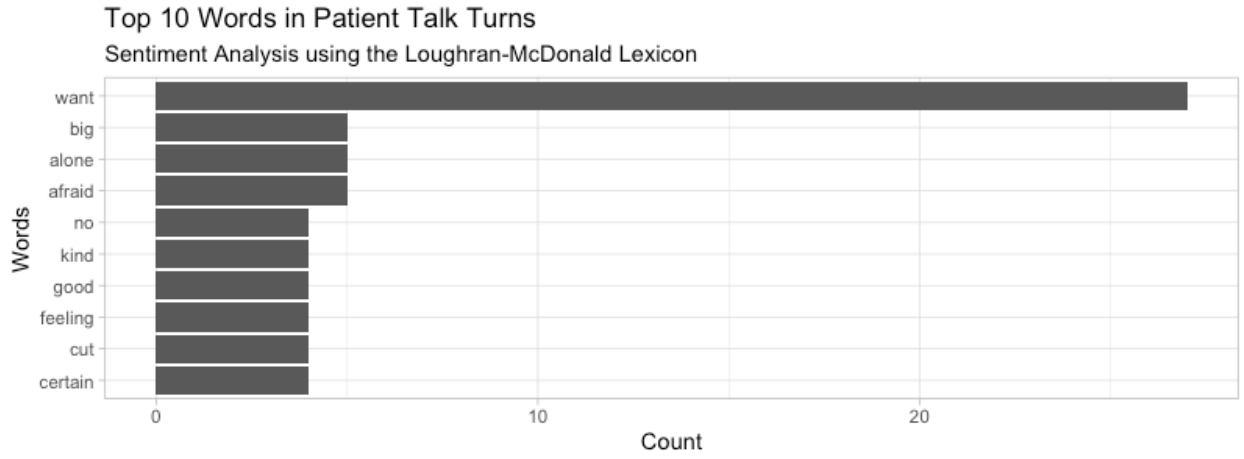

Figure 20: Top 10 words in patient talk turns by Loughran-McDonald lexicon

The sentiment analysis performed by the AFINN lexicon (Figure 21) has identified far more positive talk turns (97) than the human raters. In fact, it also has rated a couple of talk turns more positively than humans, with two instances in minute 25 and 26 having been rated 4. In contrast to the human ratings, no neutral ratings were assessed by the lexicon. In addition, the count of all negative ratings assessed by AFINN (67) is lower than the those ones rated by humans across all talk turns (75). The lexicon also seems to identify that the session ended on a positive note, having four consecutive positive ratings from minute 48 on. Just like human raters, it also identified a prevalence of negative ratings in the beginning of the session.

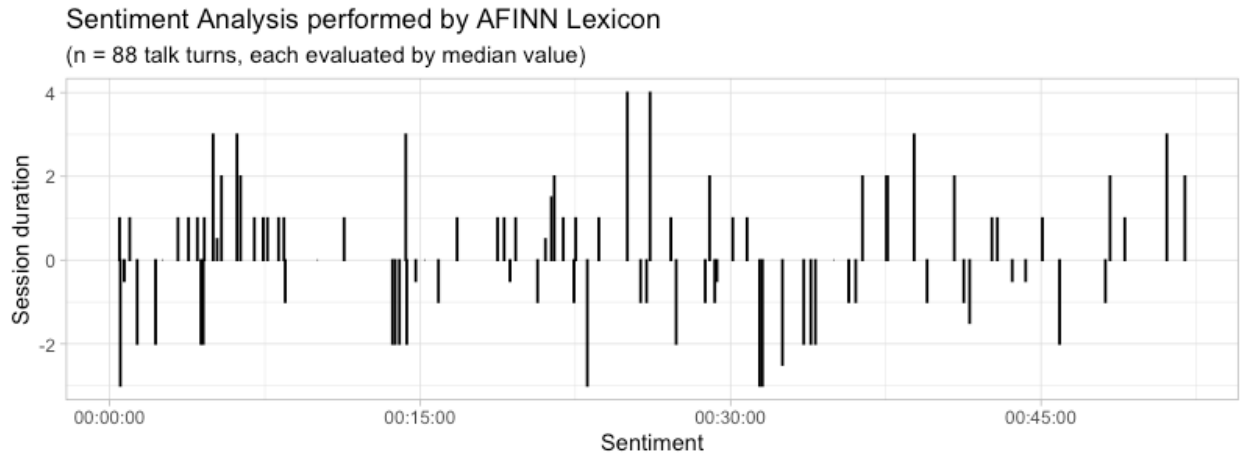

Figure 21: Sentiment Analysis performed by AFINN lexicon

Assessing Inter-Rater Reliability for the Human Raters and AFINN Lexicon

Human raters intra-rater reliability for the ordinal data was calculated by applying Krippendorff's alpha and an intra-class correlation ICC(2,1). Krippendorff's alpha evaluated both the total number of raters with missing ratings and a subset of raters with complete ratings for all talk turns, with a confidence level = 0.95, and 100 bootstrap samples. The alpha value for 52 raters and 180 observations/talk turns was $\alpha = 0.440$, CI = 0.39 to 0.47. For the subset of raters n = 41 with same number of observation the alpha value was $\alpha = 0.444$, CI = 0.39 to 0.48. The intra-class correlation value for agreement between human raters was ICC = 0.413, CI = 0.36 to 0.47, confidence level = 0.95. Therefore, Krippendorff's alpha for human raters denotes low reliability, while the ICC calculation indicates the absolute agreement between raters is fair.

For the evaluation of the performance of the AFINN lexicon, we used Krippendorff's alpha, with a confidence level = 0.95, and an intra-class correlation ICC(3,1) to measure consistency between the lexicon and human raters. Both Krippendorff's alpha for 176 and 88 observations retrieved the same alpha value $\alpha = 0.339$, with similar confidence interval values: CI = 0.13 to 0.47 for 176 observations and CI = 0.11 to 0.50 for 88 observations. The ICC(3,1) measurement for the 188 observation indicates an ICC = 0.49, with a CI = 0.31 to 0.63. These results suggest that there is a low reliability according to Krippendorff's alpha and a fair reliability in terms of consistency between human raters and the AFINN lexicon.

5. Discussion

We set out to investigate the application of sentiment analysis in a lexicon-based approach to psychotherapy session data. We were interested in not only using actual in-person psychotherapy data but also to evaluate the reliability of using existing general-purpose lexicons for the analysis. We used four lexicons available in the *tidytext* package (Silge & Robinson, 2016) to perform the analysis of a full psychotherapy session.

First, we tested the performance of the NRC lexicon, evaluating how well it performs in identifying emotion in text by comparing it to a sample of human raters. The human raters' inter-rater reliability was measured by assessing Krippendorff's alpha and Fleiss' kappa. The level of agreement among human raters was unreliable according to the Krippendorff and fair according to Fleiss. Even though the evaluation of the level of agreement between human raters and the NRC

was considered unreliable by Krippendorff's and fair by Cohen's kappa standards it did obtain a higher level of agreement than the group of human raters. This could indicate that the average of human raters performs better than individual human raters or that the NRC lexicon does have the ability to moderately identify emotion from psychological text.

There are some aspects of emotion detection that should be addressed regarding the way human raters identified emotion in text in comparison to how the lexicon achieved it. Human raters were prompted to identify the prevalent emotion in each talk turn, while the lexicon finds terms in text and matches them to pre-labeled emotions. Due to the fact that we cannot compute inter-rater reliability between multiple answers to the same observation by the same rater, our methodology design used an average to determine the overall emotion present in the patient's speech. That allowed us to assess whether or not the lexicon was competent enough to identify that same emotion in the same fragment of text. While we had anticipated and planned for this challenge in our methodology design, the detection of emotion utilizing this technique or this particular lexicon does not seem to meet a proper standard of efficiency or reliability which would allow it to be applied outside the scope of a research study. In the context of a psychotherapy session it is possible for a patient to express diverse and complex emotions throughout different moments in time. Yet, it does not seem very plausible that contradicting, unrelated, or numerous emotions are being expressed simultaneously at each talk turn. On the other hand, the list of emotions covered in the NRC lexicon, based on Plutchik's (2001) model representation of emotions, does not seem to be adequate to fully capture the nuance nor the range of emotions expressed in a clinical setting.

The analysis of sentiment using the three lexicons shares some of NRC's shortcomings. The short range of possible ratings (positive or negative) in the nominal sentiment lexicons, Bing and Loughran-McDonald, was expected to contribute to a higher level of agreement between the lexicons and the average human ratings. This did not occur. Human raters' best agreement scores for both lexicons was Krippendorff's alpha $\alpha = 0.34$ and Fleiss' kappa $\kappa = 0.34$. Rater group size and observations/talk turns evaluated generated different confidence interval values but the average alpha and kappa were technically equal. The assessment of agreement between human ratings and the nominal lexicons also generated similar scores according to Krippendorff's alpha and Cohen's kappa: for Bing it generated $\alpha = 0.48$ and $\kappa = .48$; and for Loughran-McDonald $\alpha = 0.52$ and $\kappa = 0.51$. While for Krippendorff this indicated unreliability because values are under 0.60, for Cohen it revealed fair and moderate agreement between raters and the lexicons.

The AFINN lexicon was the exception to the three other lexicons' difference in ratings between one group of raters and the other. It was the only instance where Krippendorff's alpha value had a higher agreement value for human raters ($\alpha = 0.44$) than what it calculated for the agreement between human raters and lexicon ($\alpha = 0.34$). The culprit for this discrepancy might lie in the fact that human raters rated sentiment as neutral across many talk turns, while no neutral ratings were assessed by the lexicon. On the other hand, the higher ICC value for the agreement between human ratings and AFINN ratings (ICC = 0.49) in comparison to the ICC agreement between human raters (ICC = 0.41) might be related to the fact that, while ICC(3,1) measures the magnitude between rating values, ICC(2,1) measures agreement based on absolute ratings, not magnitude. In addition, we also question the reliability of the ratings associated with the terms in the AFINN lexicon.

Several factors might also have negatively influenced agreement between human raters. One of the issues might be related to respondent fatigue. Feedback provided by some respondents mentioned that evaluation of a complete psychotherapy session was lengthy; some respondents pointed out that the text was "sad" and "boring"; and that the text didn't make sense. The fact that some respondents only filled out the first questionnaire and not the second might be in part be to account of those issues. Another factor which might also have played a role in the low agreement scores might have been due to the fact that participants were not trained raters and might not have had previous experience in coding for sentiment or emotion. Also, even though the sample consisted of more native English speakers than those who speak English as a second language, the inclusion of the latter might have skewed the results.

While the inter-rater agreement scores are inconclusive and indicate fair to moderate agreement levels, it is interesting to note that we did not observe any scores in our analyses that indicated disagreement. Negative values in Krippendorff and other IRR allow for measurement of disagreement and none of the agreement scores were equal or inferior to zero. Further research is necessary to validate the four lexicons used. We would recommend focusing on the development of psychotherapy specific-purpose lexicons, which would allow for more consistent ratings of sentiment and appropriate detection of emotion. The exploration of the application of the alternate sentiment analysis technique, the text classification approach, might also inform the development of more appropriate and reliable methods to analyze sentiment, emotion, and content in psychotherapy sessions.

Bibliography

Al-Mosaiwi, M., & Johnstone, T. (2018). In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation. *Clinical Psychological Science, 6*(4), 529-542.

Althoff, T., Clark, K., & Leskovec, J. (2016). Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. In L. Lee (Ed.), *Transactions of the Association for Computational Linguistics. 4*, pp. 463-476. Association for Computational Linguistics.

Bing, L. (2019). *Opinion Mining, Sentiment Analysis, and Opinion Spam Detection*. Retrieved from https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

Boyd, R. L. (2017). Psychological Text Analysis in the Digital Humanities. In S. Hai-Jew, *Data Analytics in Digital Humanities.* Cham, Switzerland: Springer International Publishing.

Boyd, R. L., & Pennebaker, J. W. (2016). A Way With Words: Using Language for Psychological Science in the Modern Era. In C. V. Dimofte, C. P. Haugtvedt, & R. F. Yalch, *Consumer psychology in a social media world* (pp. 222-236). New York, NY, USA: Routledge/Taylor & Francis Group.

Brooks, H. (1994, September). The relationship between science and technology. *Research Policy, 23*(5), 477-486.

Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural Language Processing in mental health applications using non-clinical texts. *Natural Language Engineering, 23*(5), 649-685.

Can, D., Marín, R. A., Georgiou, P. G., Imel, Z. E., Atkins, D. C., & Narayanan, S. S. (2016). "It sounds like...": A Natural Language Processing Approach to Detecting Counselor Reflections in Motivational Interviewing. *Journal of Counseling Psychology, 63*(3), 343-350.

Dalianis, H. (2018). *Clinical Text Mining. Secondary Use of Electronic Patient Records.* Cham, Switzerland: Springer International Publishing.

Davcheva, E. (2018). Text Mining Mental Health Forums - Learning from User Experiences. *Twenty-Sixth European Conference on Information Systems (ECIS 2018)*, (pp. 1-11). Portsmouth, UK.

Fosha, D. (2009). Emotion and Recognition at Work: Energy, Vitality, Pleasure, Truth, Desire & The Emergent Phenomenology of Transformational Experience. In D. Fosha, D. J. Siegel, & M. F. Solomon, *The Healting Power of Emotions. Affective Neuroscience, Development & Clinical Practice* (pp. 172-203). New York, NY, USA: W. W. Norton & Company, Inc.

Freud, S. (1900/1913). *The Interpretation of Dreams.* New York, NY, US: MacMillan Company.

Freud, S. (1914/1990). *The Psychopathology of Everyday Life.* New York, NY, UK: W. W. Norton & Company.

Freud, S. (1973/1991). *Introductory Lectures on Psychoanalysis.* Middlesex, UK: Penguin Books.

Gamer, M., Lemon, J., & Singh, I. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement.* Retrieved from https://CRAN.R-project.org/package=irr

Gaut, G., Steyvers, M., Imel, Z. E., Atkins, D. C., & Smyth, P. (2017). Content Coding of Psychotherapy Transcripts Using Labeled Topic Models. *IEEE Journal of Biomedical and Health Informatics, 21*(2), 1-12.

Goldberg, S. B., Rousmaniere, T., Miller, S. D., Whipple, J., Nielsen, S. L., Hoyt, W. T., & Wampold, B. E. (2016). Do Psychotherapists Improve with Time and Experience? A Longitudinal Analysis of Outcomes in a Clinical Setting. *Journal of Counseling Psychology, 63*(1), 1-11.

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23-34.

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using Smartphones to Collect Behavioral Data in Psychological Science: Opportunities, Practical Considerations, and Challenges. *Perspectives on Psychological Science, 11*(6), 838-854.

Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures, 1*(1), 77-89.

He, Q. (2018, January --). *Text Mining and Item Response Theory for Psychiatric and Psychological Assessment*. Retrieved July 2019, from APA Div. 5: Quantitative and Qualitative Methods: https://www.apadivisions.org/division-5/publications/score/2018/01/text-mining

Hirsch, T., Soma, C., Merced, K., Kuo, P., Dembe, A., Caperton, D. D., . . . Imel, Z. E. (2018). "It's hard to argue with a computer:" Investigating Psychotherapists' Attitudes towards Automated Evaluation. *DIS (Des Interact Syst Conf)*, 559-571.

Holmes, E. A., Ghaderi, A., Harmer, C. J., Ramchandani, P. G., Cuijpers, P., Morrison, A. P., . . . Craske, M. G. (2018). The Lancet Psychiatry Commission on psychological treatments research in tomorrow's science. *The Lancet Psychiatry, 5*(3), 237-286.

Iliev, R., Dehghani, M., & Sagi, E. (2015, June). Automated text analysis in psychology: methods, applications, and future developments. *Language and Cognition, 7*(2), 265-290.

Imel, Z. E., Caperton, D. D., Tanana, M., & Atkins, D. C. (2017). Technology-enhanced Human Interaction in Psychotherapy. *Journal of Counseling Psychology, 64*(4), 385-393.

Imel, Z. E., Steyvers, M., & Atkins, D. C. (2015). Computational Psychotherapy Research: Scaling up the Evaluation of Patient-Provider Interactions. *Psychotherapy, 52*(1), 19-30.

Krippendorff, K. (2004). *Content Analysis: An Introduction to its Methodology.* Thousand Oaks, CA, USA: Safe Publications Inc.

Liu, B., & Zhang, L. (2012). A Survey of Opinion Mining and Sentiment Analysis. In C. Aggarwal, & C. Zhai, *Mining Text Data.* Boston, MA, USA: Springer.

Loughran, T., & McDonald, B. (2011). *Software Repository for Accounting and Finance (University of Notre Dame)*. Retrieved from https://sraf.nd.edu/textual-analysis/resources/

Martinez, V. R., Flemotomos, N., Ardulov, V., Somandepalli, K., Goldberg, S. B., Imel, Z. E., . . . Narayanan, S. (2019, September 15-19). Identifying Therapist and Client Personae for Therapeutic Alliance Estimation. *Interspeech*, 1901-1905.

Martire, R. L. (2017). *rel: Reliability Coefficients.* Retrieved from ttps://CRAN.R-project.org/package=rel

Melina, A. (1997). *Computer-assisted text analysis methodology in the social sciences.* Mannheim: Zentrum für Umfragen, Methoden und Analysen -ZUMA-.

Mergenthaler, E., & Stinson, C. (1992). Psychotherapy Transcription Standards. *Psychotherapy Research*, 125-142.

Mieskes, M., & Stiegelmayr, A. (2018). Preparing Data from Psychotherapy for Natural Language Processing. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (pp. 2896-2902). LREC.

Mohammad, S. M. (2006). *NRC Word-Emotion Association Lexicon (aka EmoLex)*. Retrieved July 2019, from https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

Montag, C., Duke, É., & Markowetz, A. (2016). Toward Psychoinformatics: Computer Science Meets Psychology. *Computational and Mathematical Methods in Medicine, Vol. 2016*(Article ID 2983685), 1-10.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying Words: Predicting Deception From Linguistic Styles. *Personality and Social Psychology Bulletin, 29*(5), 665-675.

Nielsen, F. A. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*(arXiv:1103.2903), 93-98.

Owen, J., & Imel, Z. E. (2016). Introduction to the Special Section "Big'er Data": Scaling Up Psychotherapy Research in Counseling Psychology. *Journal of Counseling Psychology, 63*(3), 247-248.

Pace, B. T., Xiao, B., Aaron, D., Steyvers, M., Narayanan, S., Atkins, D. C., & Imel, Z. E. (2016). What About the Words? Natural Language Processing in Psychotherapy. *Psychotherapy Bulletin*, 14-18.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic Inquiry and Word Count (LIWC): A Computer-Based Text Analysis Program.* Mahwah, NJ, USA: Erlbaum Publishers.

Plutchik, R. (2001). The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist, 89*(4), 344-350.

Provoost, S., Ruwaard, J., van Breda, W., Riper, H., & Bosse, T. (2019). Validating Automated Sentiment Analysis of Online Cognitive Behavioral Therapy Patient Texts: An Exploratory Study. *Frontiers in Psychology, 10*, 1-12.

Quercia, D., Kosinski, M., Stillwell, D., & Crowcroft, J. (2011). Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. *2011 IEEE Third International Conference*

on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, 180-185.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Retrieved from https://www.R-project.org/

Reshma, A. J., James, J. J., Kavya, M., & Saravanan, M. (2016). An Overview of Character Recognition Focused on Offline Handwriting. *ARPN Journal of Engineering and Applied Sciences, 11*(15), 9372-9378.

Revelle, W. (2018). *psych: Procedures for Psychological, Psychometric, and Personality Research.* Retrieved from https://CRAN.R-project.org/package=psych

Rezaii, N., Walker, E., & Wolff, P. (2019). A machine learning approach to predicting psychosis using semantic density and latent content analysis. *npj Schizophrenia, 5*(9), 1-12.

Santilli, S., & Nota, L. (2017). The use of Latent Semantic Analysis in the Positive Psychology: a Comparison with Twitter Posts. *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, (pp. 494-498).

Schroder, T., Orlinsky, D., Ronnestad, M. H., & Willutzki, U. (2015). Psychotherapeutic Process from the Psychotherapist's Perspective. In O. C. Gelo, A. Pritz, & B. Rieken, *Psychotherapy research: Foundations, process, and outcome* (pp. 351-365). New York, NY, US: Springer-Verlag Publishing.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin, 86*(2), 420-428.

Silge, J., & Robinson, D. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *The Open Journal, 1*(3), http://dx.doi.org/10.21105/joss.00037. Retrieved from http://dx.doi.org/10.21105/joss.00037

Silge, J., & Robinson, D. (2017). *Text Mining with R. A Tidy Approach.* Sebastopol, CA, USA: O'Reilly Media, Inc.

Skovholt, T. M., & Ronnestad, M. H. (2003). Struggles of the Novice Counselor and Therapist. *Journal of Career Development, 30*(1), 45-58.

Steiner, M. D., Phillips, N. D., & Trutmann, K. (2019, June 2). *ShinyPsych: An easy way to program psychology experiments using Shiny*. Retrieved from rdrr.io: https://rdrr.io/github/ndphillips/ShinyPsych/

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics, 37*(2), 267-307.

Tanana, M. J., Soma, C. S., Srikumar, V., Atkins, D. C., & Imel, Z. E. (2019). Development and Evaluation of ClientBot: Patient-Like Conversational Agent to Train Basic Counseling Skills. *Journal of Medical Internet Research, 21*(7), 1-13.

Tanana, M., Hallgreen, K., Atkins, D., Smyth, P., & Srikumar, V. (2015). Recursive Neural Networks for Coding Therapist and Patient Behavior in Motivational Interviewing. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 71-79.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology, 29*(1), 24-54.

Vinyals, O., & Le, Q. (2015, July 22). *A Neural Conversational Model.* Retrieved August 2019, from arXiv.org: https://arxiv.org/abs/1506.05869

Wang, S.-H., Ding, Y., Zhao, W., Huang, Y.-H., Perkins, R., Zou, W., & Chen, J. J. (2016). Text Mining for Identifying Topics in the Literatures about Adolescent substance use and depression. *Public Health, 16*, 1-8.

Wickham, H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse' (R Package version 1.2.1.)*. Retrieved from https://CRAN.R-project.org/package=tidyverse

Xiao, B., Imel, Z. E., Georgiou, P. G., Atkins, D. C., & Narayanan, S. S. (2015). "Rate My Therapist": Automated Detection of Empathy in Drug and Alcohol Counseling via Speech and Language Processing. *PLOS ONE, 10*(12: e0143055), 1-15.

Yarkoni, T. (2012, December). Psychoinformatics: New Horizons at the Interface of the Psychological and Computing Sciences. *Current Directions in Psychological Science, 21*(6), 391-397.

Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., & Manning, J. R. (2018). Is Automatic Speech-to-text Transcription Ready for Use in Psychological Experiments? *Behavior Research Methods, 50*, 2597-2605.