Analyzing the Quality of Counseling Conversations: the Tell-Tale Signs of High-quality Counseling

Verónica Pérez-Rosas¹, Xuetong Sun¹, Christy Li¹, Yuchen Wang¹
Kenneth Resnicow² and Rada Mihalcea¹
¹Computer Science and Engineering, ²School of Public Health
University of Michigan
{vrncapr, xuetong, christyq, lucasyc, kresnic, mihalcea}@umich.edu

Abstract
Behavioral and mental health are pressing issues worldwide. Counseling is emerging as a core treatment for a variety of mental and behavioral health disorders. Seeking to improve the understanding of counseling practice, researchers have started to explore Natural Language Processing approaches to analyze the nature of counseling interactions by studying aspects such as mirroring, empathy, and reflective listening. A challenging aspect of this task is the lack of psychotherapy corpora. In this paper, we introduce a new dataset of high-quality and low-quality counseling conversations collected from public web sources. We present a detailed description of the dataset collection process, including preprocessing, transcription, and the annotation of two counseling micro-skills: reflective listening and questions. We show that the obtained dataset can be used to build text-based classifiers able to predict the overall quality of a counseling conversation and provide insights into the linguistic differences between low-quality and high-quality counseling.

Keywords: conversation analysis, counseling, mental health

1. Introduction
Mental and behavioral disorders, such as substance abuse, are top on the list of the most costly and prevalent conditions worldwide.

Particularly in the US, a recent survey on public health reported that in 2014 3.3% of all adults had co-occurring mental illness and substance abuse disorders.

As behavioral counseling has been shown to be an effective treatment method for these conditions, the number of people seeking counseling services is increasing (Chava, 2014). Despite its potential benefits, such as combating addiction and providing broader disease prevention and management, the mechanisms behind successful behavioral counseling have not been fully elucidated (Moyers et al., 2009).

Specific counseling skills have shown to increase the likelihood of positive health outcomes (Güme et al., 2009; Vader et al., 2010). Regardless of the counseling method, counselors follow general principles, such as supporting autonomy, expressing empathy, centering on the patient and engaging patients using specific skills such as reflective listening (Charles et al., 1997; Harting et al., 2004). In contrast, using a more directing style – characterized by counselors providing instruction and advice, and patients obeying, adhering and complying (Miller and Rollnick, 2013) – is usually avoided.

The guidelines described above can be used to differentiate between low and high quality counseling. Thus, in a broad classification, psychotherapy conversations where counselors follow preferred practices can be considered as high-quality (or guideline adhering) counseling, whereas those conversations where they do not can be regarded as low-quality counseling (or guideline non-adhering).

Following this idea, our paper analyzes counseling conversations with the final goal of distinguishing between low and high quality counseling. In particular, we focus our analysis on counseling conducted using Motivational Interviewing (MI), a well-established evidence-based counseling style for treating addiction and other behaviors (Moyers et al., 2009; Catley et al., 2012; Apodaca et al., 2014).

Our work makes two main contributions. First, we introduce a new dataset of counseling conversations collected from public web sources. With this dataset, we seek to address the problem of lack of psychotherapy corpora for NLP applications, as most of current psychotherapy corpora have important constrains regarding their public accessibility due to ethical and privacy concerns. Second, we show that the collected dataset can be used to build text-based classifiers able to predict the overall quality of a counseling conversation and provide insights into the tell-tale signs of high-quality counseling.

2. Related Work
While clinical mental health counseling has been found useful in the treatment of public health issues, evaluating its quality remains a problem. This is mainly because most studies on clinical psychology have been limited by the need for human-based evaluation and by small sample sizes.

Computational approaches for the analysis of counseling interactions have focused on two main lines of work.

First, seeking to develop tools for the automatic evaluation of counseling practice, several linguistic based approaches have been proposed to aid the automatic identification of counselor and client behaviors that are correlated to successful interventions (Klonek et al., 2015). Can et al., 2012 used n-grams, similarity features between coun-
selor and client speech, and dialog meta-features to automatically detect and code counselors’ reflective listening. A method based on labeled topic models is presented in Atkins et al., 2012, Atkins et al., 2014, where authors focus on automatically identifying conversation topics that relate to counselor behaviors such as reflective listening, questions, support, and empathy. Methods that combine acoustic and linguistic datastreams have also been proposed to evaluate the quality of counseling interactions. Xiao et al., 2014 presented a study on the automatic evaluation of counselor empathy based on analyzing correlations between prosody patterns and empathy showed by the therapist during counseling interactions.

Second, aiming to improve the understanding of counseling interactions, researchers have started to explore NLP approaches to study aspects such as language mirroring, empathy, and reflective listening. Tanana et al., 2015 addressed the identification of counselor’s statements that discuss client’s change talk using recursive neural networks to model sequences of counselor and client verbal exchanges. Lord et al., 2015 analyzed the language style synchrony between therapists and clients during MI encounters. Their approach relies on the psycholinguistic categories from the Linguistic Inquiry and Word Count lexicon to measure the degree to which the counselor language matches the client language. More recently, Althoff et al., 2016 explored language style and symmetry in counseling interactions by analyzing a large sample of text-message-based counseling. Their main findings suggest that counselors who are more successful act with more control in the conversations and show lower levels of verbal coordination (mirroring) than their less successful counterparts.

Furthermore, there are ongoing efforts on creating annotated resources that facilitate NLP advances in the analysis of clinical text in applications such as automatic annotation of pathology reports and oncology reports as well as data from biomedical journals Roberts et al., 2007, Albright et al., 2013 Verspoor et al., 2012. Despite this efforts, to our knowledge, there are only few psychotherapy corpora available. One of them is the “Alexander Street Press”, which is a large collection of transcripts and video recordings of therapy sessions on different subjects such as anxiety, depression, family conflicts, and others. There are also other psychology datasets available under limited access from the National Institute of Mental Health (NIMH).

In this paper, we present the development of a counseling conversations dataset that can be used to implement data-driven methods for the automatic evaluation of counseling quality. We specifically focus on the overall conversation quality, with the final goal of providing linguistic cues associated with high-quality counseling.

4http://psychiatry.yale.edu/pdc/resources/datasets.aspx

3. A Dataset of Low and High Quality Counseling

3.1 Collecting Counseling Conversations from the Web

We started by identifying video clips containing brief counseling interactions conducted using Motivational Interviewing (MI) from publicly available video-sharing sources such as YouTube and Vimeo. Keywords used to search for these videos include “motivational interviewing”, “MI counseling”, “effective MI”, “good MI”, “MI counseling demonstration”, “role play MI” for the high-quality category, and “ineffective MI”, “bad MI”, “bad counseling”, “how not to do MI”, “the bad counselor” for the low-quality category. To select the videos, we used the following guidelines: the video should include only two participants, i.e., counselor and client; the video should include minimal interruptions, such as background narrative, music, or animation; the session should address a behavior change e.g. smoking cessation, drinking; and finally, the counselor-client interaction should last at least 3 minutes.

The obtained recordings consist mainly of MI counseling demonstrations from MI training services and students’ MI role-play practice from undergraduate-level psychology courses. The sessions address various health topics including smoking cessation, alcohol consumption, substance abuse, weight management, mental disorders, and medication adherence and portray several practice settings such as private practice, school counseling, and pharmacy counseling among others.

After collecting our initial pool of videos, we conduct a second filtering step to verify that the counseling was conducted using MI and that the video caption matched the video content, i.e., portray either a high-quality or a low-quality counseling interaction. To evaluate MI use (or the lack of it) we followed the guidelines in MI literature Miller and Rollnick, 2013. The criteria to label a counseling interaction as either low or high quality is as follows: during high-quality counseling, the conversation should present, to some extent, reflective listening, questions, as well as collaboration and support. In contrast, the low-quality counseling should show a predominant directive style, which includes confrontation, advising without permission, and lack of listening.

The final video set includes 151 counseling conversations. From this, 72 video clips were labeled as high-quality counseling and the remaining 79 as low-quality counseling. The length of the conversations varies from 5-20 minutes. Table shows transcript excerpts corresponding to high-quality and low-quality counseling conversations in the dataset.

Preprocessing. All the videos are first converted into standard mp4 format and then preprocessed to address issues frequently present in shared video content such as introductory titles, animations, and narratives. In most cases these interruptions appeared only at the beginning of the video so we manually trimmed that portion of the video until the counselor-patient interaction started. This process can also be optimized using automatic methods such as optical character and facial recognition, however, we opted for a manual approach in order to obtain accurate examples.
Table 1: Transcript excerpts corresponding to high-quality and low-quality counseling conversations

<table>
<thead>
<tr>
<th>Code</th>
<th>Count</th>
<th>Verbal examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question</td>
<td>1122</td>
<td>What do you think it would take to change your mind about participating in physical activity?</td>
</tr>
<tr>
<td>Reflection</td>
<td>813</td>
<td>It sounds like you're concerned by your weight and you want to start to make positive changes.</td>
</tr>
</tbody>
</table>

Table 2: Frequency counts and verbal examples of Questions and Reflections in the dataset

of counselor-patient interactions.

**Transcription.** In order to transcribe the video clips, we adopt a semi-automatic approach. First, we use the YouTube automatic captioning to obtain the conversation transcript and then we manually labeled the conversation turns as either counselor or patient speech.

### 3.2. Annotation of Counseling Skills

Seeking to evaluate the counseling interaction between counselors and patients, we decided to annotate two core interview micro-skills in counseling practice: reflective listening and questions (Tollison et al., 2008). These two micro-skills are assessed using the Motivational Interviewing Treatment Integrity (MITI) coding scheme version 4 (Moyers et al., 2016), which is the current gold standard for MI fidelity evaluation. Hence, all the video clips in our dataset are manually annotated to identify questioning (Questions) and reflective listening statements (Reflections).

In order to conduct the annotation, two undergraduate students were trained in the use of the MITI 4.0. During this training, the annotators learned how to parse the counselor-patient interaction (i.e., deciding which portion of the conversation shows the given behavior), practiced the correct assignment of behavior codes, and conducted team coding on sample sessions.

The 151 sessions were randomly distributed among the two annotators. During the coding process, the annotators used both the audio recording and the transcript. The annotation was conducted at conversation turn-level using Nvivo 5, an annotation and quantitative analysis suite that allows selecting text in the transcript and labeling it with a given code, e.g., reflection or question.

In order to verify the reliability of the annotations, we calculated the inter-annotator agreement in a sample of 20 counseling conversations, with even distribution for the Low-quality and High-quality categories. The intra-class correlation scores for both Questions and Reflections codes are 0.96 and 0.94, respectively, thus showing good levels of agreement between the two annotators.

### 4. Discriminating Between High-quality and Low-quality Counseling

#### 4.1. Analysis of Counseling Conversations

We start by exploring linguistic differences between the counseling interactions to get insights into the mechanisms of high-quality counseling. Our analyses are based on the semantic word classes from the LIWC lexicon and the semantic word-class scoring by (Mihalcea and Pulman, 2009). Table 3 shows the top classes for both, low and high quality counseling.

The results show interesting differences between the two types of conversations. While high-quality counseling focuses on aspects related to encouragement and reflective listening such as family, positive feeling, feelings, and hearing, low-quality counseling shows a more directive lan-
4.2. Distinguishing Between High and Low Quality Counseling

In this section, we explore the use of linguistic cues to build a computational model that predicts the overall quality of counseling. This could provide information on whether counselors focus on positive or negative aspects of the client communication, and how this relates to the conversation quality, i.e., low or high. Thus, we analyze the sentiment expressed by counselors during each turn in the conversation. Given the effort required to manually annotated the sentiment in each conversation, we opted for using an automatic off-the-shelf sentiment classifier from the Stanford Core NLP package (Manning et al., 2014). We obtain a sentiment score for each counselor turn, scored from very negative to very positive, and calculate the percentage of positive, negative, and neutral turns during the conversation. Figure 2 shows the sentiment distribution over the high-quality and low-quality conversations. The box plots in the figure suggest differences between the low and high quality groups, particularly for positive sentiment. In order to look more closely into the positive sentiment trend during the counseling encounters, we plot the distribution of positive turns by the counselor across the low and high quality counseling conversation. The plot in Figure 3 shows that counselors increasingly focus on positive aspects of the client expressions, particularly during high-quality conversations.

Table 3: Results from LIWC word class analysis. Top ranked semantic classes associated to low and high quality counseling are shown.

<table>
<thead>
<tr>
<th>Class</th>
<th>Score</th>
<th>Sample words</th>
<th>Class</th>
<th>Score</th>
<th>Sample words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family</td>
<td>2.04</td>
<td>Mom, wife, parents, husband</td>
<td>Self</td>
<td>1.33</td>
<td>I, we, me, my, our</td>
</tr>
<tr>
<td>Feel</td>
<td>1.85</td>
<td>Feel, pain, feeling, sense</td>
<td>Negate</td>
<td>1.31</td>
<td>Not, don’t, no, can’t, without</td>
</tr>
<tr>
<td>Posfeel</td>
<td>1.52</td>
<td>Like, care, enjoy, glad</td>
<td>Inhibition</td>
<td>1.81</td>
<td>quit, stop, control, avoid</td>
</tr>
<tr>
<td>Anxiety</td>
<td>1.38</td>
<td>Afraid, worried, overwhelmed</td>
<td>Time</td>
<td>1.06</td>
<td>Now, start, today, before</td>
</tr>
<tr>
<td>Optimism</td>
<td>1.31</td>
<td>Ready, hope, confidence, determined</td>
<td>Present</td>
<td>1.06</td>
<td>Know, do, need, want</td>
</tr>
<tr>
<td>Hear</td>
<td>1.25</td>
<td>Sounds, heard, talking, said</td>
<td>Pronoun</td>
<td>1.04</td>
<td>You, I, it, your, we</td>
</tr>
</tbody>
</table>

Figure 1: Distribution of reflections, questions, listening (hear), optimism, family and negation (negate) word classes in high and low quality counseling

Figure 2: Distribution of positive, negative, and neutral counselor turns in high and low quality counseling

Figure 3: Distribution of positive turns by the counselor across low and high quality counseling
the counseling conversation

The feature set consists of the cues identified during our exploratory analyses as potential indicators of counseling quality, as well as additional text features used during standard NLP feature extraction such as ngrams. The features are extracted from the transcripts of the counseling conversations. During our experiments, we first explore the predictive power of each cue separately, followed by an integrated model that attempts to combine all the linguistic cues to improve the prediction of counseling quality. The different features are as follows:

**N-grams:** These features represent the language used by the counseling-conversation participants and include all the unique words and word-pairs present in the transcript. We extract a vector containing the frequencies of each word and word pair present in the transcript.

**Semantic information:** We use categories from the LIWC (Tausczik and Pennebaker, 2010), Opinion Finder (Wilson et al., 2005) and the Wordnet Affect (Strapparava and Valitutti, 2004) lexicons to derive features that identify identifying words as belonging to certain semantic categories that are potential markers of the conversation quality.

**Metafeatures:** We also extract a set of metafeatures that describe the conversation interaction, including the number of counselor turns, client turns, average words during client and counselor turns, and the ratio of counselor and client words in each turn.

**Sentiment:** These features are designed to capture the sentiment trend in the counselor responses during the conversation. To derive these features, we first obtain the sentiment expressed by counselors during each turn, scored from very negative to very positive (−−, −, 0, +, ++) using the sentiment analysis classifier from Stanford Core NLP, and then obtained a set of descriptors that capture the sentiment trend. The set includes the percentage of positive, negative, and neutral turns during the conversation, the number of times the sentiment changes during the conversation, as well as counts of sequences increasing and decreasing sentiment intensity i.e., −+, −−, +−, +++.  

**MITI behaviors** This set includes the number of reflections and questions by the counselor during the conversation as well as the ratio of reflections to questions. The counts are derived from the turn-level annotations described in section 3.2.

We conduct several experiments to discriminate between low-quality and high-quality encounters. During our experiments, the evaluations are done at conversation level. The classifiers are built using the Support Vector Machine algorithm and the different sets of linguistic features. We perform leave-one-out cross-validation in all our experiments and we use the majority class baseline as a reference value.

Results shown in Table 4 show that all the feature sets perform above the baseline, with the MITI behaviors being the best performing features, followed by the n-grams features. We also observe that the combination of all feature sets provides the best performance.

Seeking to explore the role played by the different feature sets, we conduct an ablation study, where we remove one feature set at the time from the best performing model i.e., “all features”. As observed in Table 5 the ngrams features contribute the most to the final model, followed by the lexicon features. Interestingly, the results show that the combination of n-grams and lexicons offer similar performance as the MITI behaviors features. These results are encouraging as they suggest that standard linguistic features can achieve similar performance as manually coded features (MITI behaviors) while evaluating the overall quality of counseling conversations.

### Table 4: Overall prediction results and F-scores for high-quality and low-quality counseling conversations using several linguistic feature sets.

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Acc.</th>
<th>F-score High</th>
<th>F-score Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>52.31%</td>
<td>0.82</td>
<td>0.83</td>
</tr>
<tr>
<td>Ngrams</td>
<td>82.78%</td>
<td>0.82</td>
<td>0.83</td>
</tr>
<tr>
<td>Lexicons</td>
<td>72.84%</td>
<td>0.71</td>
<td>0.73</td>
</tr>
<tr>
<td>Metafeatures</td>
<td>76.15%</td>
<td>0.77</td>
<td>0.74</td>
</tr>
<tr>
<td>MITI Behav</td>
<td>83.44%</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>Sentiment</td>
<td>70.86%</td>
<td>0.69</td>
<td>0.81</td>
</tr>
<tr>
<td>All features</td>
<td>87.41%</td>
<td>0.87</td>
<td>0.87</td>
</tr>
</tbody>
</table>

### Table 5: Feature ablation study.

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>All features</td>
<td>87.41%</td>
</tr>
<tr>
<td>– Ngrams</td>
<td>83.44%</td>
</tr>
<tr>
<td>– Lexicons</td>
<td>85.43%</td>
</tr>
<tr>
<td>– Sentiment</td>
<td>86.10%</td>
</tr>
<tr>
<td>– Metafeatures</td>
<td>87.41%</td>
</tr>
<tr>
<td>– MITI Behav</td>
<td>87.41%</td>
</tr>
</tbody>
</table>

6 As implemented in the Weka library.
5. Conclusion and Future Work

In this paper, we introduced a new dataset of low-quality and high-quality counseling conversations that were collected from public sources. Through several classification experiments, we showed that such a dataset can be used to build accurate classification models able to discriminate between low-quality and high-quality counseling, with accuracy figures up to 87%.

Furthermore, we showed that standard NLP features can provide performance similar to manually coded features for this task.

We also provided insights into the linguistics markers of high-quality counseling and showed that it is characterized by positive and encouraging language.

6. References


