What Makes a Good Counselor? Learning to Distinguish between High-quality and Low-quality Counseling Conversations

Verónica Pérez-Rosas[1], **Xinyi Wu**[1], **Kenneth Resnicow**[2], **Rada Mihalcea**[1]
[1]Computer Science & Engineering, University of Michigan, USA
[2]School of Public Health, University of Michigan, USA
{vrncapr,wuxinyi,kresnic,mihalcea}@umich.edu

Abstract

The quality of a counseling intervention relies highly on the active collaboration between clients and counselors. In this paper, we explore several linguistic aspects of the collaboration process occurring during counseling conversations. Specifically, we address the differences between high-quality and low-quality counseling. Our approach examines participants' turn-by-turn interaction, their linguistic alignment, the sentiment expressed by speakers during the conversation, as well as the different topics being discussed. Our results suggest important language differences in low- and high-quality counseling, which we further use to derive linguistic features able to capture the differences between the two groups. These features are then used to build automatic classifiers that can predict counseling quality with accuracies of up to 88%.

1 Introduction

An increasing number of people are suffering from behavioral health problems, including substance abuse, smoking cessation, or eating disorders. Because behavior is something that is generally seen as changeable, behavioral counseling has emerged as an important strategy for helping people to identify and potentially change self-destructive or unhealthy behaviors (Rollnick et al., 2008).

Despite its practical benefits, such as combating addiction and providing broader disease prevention and management, the process behind successful behavioral counseling has not been fully elucidated (Moyers et al., 2009), which in turn raises the question of what makes a good counselor.

Seeking to answer this question, our paper explores differences between high- and low-quality counseling conversations. Since the quality of a counseling intervention relies highly on the active collaboration between clients and counselors (Gaume et al., 2009; Vader et al., 2010), we explore which aspects of the collaboration process occurring between counseling participants are related to counseling quality. Our categorization of counseling quality relies on general counseling principles taken from the literature on client-centered counseling (Miller and Rollnick, 2013). Thus, conversations where the counselor centers on the client and expresses empathy are considered as high-quality interactions. In contrast, conversations where the counselor mainly provides instruction and advice, and the client complies are regarded as low-quality interactions.

Our work makes three important contributions. First, we explore the use of noisy counseling data obtained from public sources for the analysis of counseling quality. Second, we conduct extensive analyses on conversational aspects such as turn-by-turn interaction, the sentiment expressed during the interaction, linguistic alignment, and salient topics during the conversation to obtain insights into what are the patterns of high-quality counseling. Hence, identifying specific conversational strategies used by good counselors. Third, we show that features derived from our analyses, along with standard Ngram features, lead to significant improvement in the classification of counseling quality.

2 Related Work

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

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 926–935
Florence, Italy, July 28 - August 2, 2019.

counselor 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, 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 counselor during the counseling interactions.

Second, aiming to improve the understanding of counseling interactions, researchers have started to explore Natural Language Processing (NLP) approaches to study aspects such as language mirroring, empathy, and reflective listening. Tanana et al. (2015) addressed the identification of counselor statements that discuss client 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 counselors and clients. Their approach relies on the psycholinguistic categories from the Linguistic Inquiry and Word Count (LIWC) lexicon (Tausczik and Pennebaker, 2010) to measure the degree in which counselors match their clients' 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 the 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.

Following this line of work, this paper presents the development of a counseling dataset that can be used to implement data-driven methods for the automatic evaluation of counseling quality. Specifically, we conduct several linguistically inspired analyses on high-quality and low-quality counseling interactions with the final goal of providing insights into conversational strategies used by good counselors.

3 Counseling Dataset from Web Sources

Most of the current work on automatically analyzing counseling interaction has been conducted on psychotherapy corpora with ethical and privacy constraints that limit their public accessibility (Can et al., 2012; Xiao et al., 2014). This, in turn, has made it difficult to replicate and improve upon previous research findings. In this paper, we address this drawback by exploring the use of counseling conversations collected from the web.

3.1 Data Collection

We start by identifying video clips containing counseling conversations from public video-sharing sources such as YouTube and Vimeo. Since the final goal of this study is to get insights into counseling quality, we followed two main strategies to search for videos that portray either high-quality or low-quality counseling. First, since the evaluation of counseling quality changes across counseling strategies (Gottheil et al., 2002), we focus on counseling conducted using motivational interviewing (MI) only. MI is a well established behavioral counseling strategy that has been successful in achieving behavioral health outcomes (Apodaca et al., 2014). Thus, we restricted our search to video titles indicating that the counseling was conducted using MI and (optionally) including information about the quality of the interaction. Specifically, we use keywords such as *effective MI*, *using MI*, *good MI*, *MI counseling demonstration*, *role play MI*, *ineffective MI*, *bad MI*, *bad MI counseling*, and *how not to do MI*. Second, to make sure that the videos portray counseling conversations we also enforce the following requirements: the video should include only two participants, i.e., counselor and client; the video should not include (or include minimal) background narratives, music, or animations; the conversation should address a behavior change, e.g., smoking cessation or quit drinking; and finally, the conversation should last at least three minutes.

After collecting our initial set of videos using the described guidelines, we conduct a second filtering step to verify that the counseling is conducted using MI and that the video caption matches the video content. To evaluate the use of MI (or the lack of it) we follow general guidelines based on the MI literature (Miller and Rollnick, 2013). The criteria to label the quality of a counseling conversation are as follows: during high-quality counseling, counselors should use (to some extent) reflective listening, ask questions,

LOW-QUALITY COUNSELING	HIGH-QUALITY COUNSELING
T: How much are you drinking? *C*: I don't think I'm drinking that much I mean it's, it's mainly for social gatherings like ... it's nothing that I do like by myself or whatever it's just that *T*: Is it like every weekend? *C*: Every other weekend I would say *T*: NAME I'm just so concerned you know can't you think of anything better to do	*T*: So, the last thing you'd want is for your daughter to start smoking *C*: I smoked when I was young and I'd certainly don't want it for her *T*: And it sounds like you're smoking setting an example those are some things that that you're also a bit concerned about but as you said earlier not, not really ready to put down your cigarettes immediately

Table 1: Transcript excerpts from low-quality and high-quality counseling conversations. *C* stands for client and *T* stands for counselor (therapist)

provide support to the client decisions, and collaborate with the client. In contrast, in low-quality conversations, counselors should show a lack of listening and a predominant directing style, with the counselor confronting the client and providing advice without asking for permission. Following these guidelines, we manually inspected all the videos. During this process, we discarded videos that did not fit our criteria. The final video set includes 259 counseling conversations, with 155 video clips labeled as high-quality counseling and 104 labeled as low-quality counseling. The length of the conversations in the dataset ranges from 5-20 minutes.

Our final set of videos consist of MI counseling demonstrations by professional counselors, and MI role-play counseling by psychology students. Each video portrays different speakers and the conversations cover various health topics including smoking cessation, alcohol consumption, substance abuse, weight management, and medication adherence. It is important to note that despite the fact that some of these conversations do not portray real patients, they are still valuable as a data source as clinical studies often use simulated patients[1] to improve the communication skills of medical practitioners (Imel et al., 2014).

3.2 Preprocessing and Transcription

All the videos in the dataset are first converted into standard mp4 format and then preprocessed to address issues frequently present in user-generated video content such as introductory titles, animations, and narratives. In most cases, these interruptions appear only at the beginning of the video so we manually trim that portion of the video until the counselor-client interaction starts. We obtain the corresponding video transcripts using YouTube automatic captioning and align it to the selected segments using the transcript time stamps. To enable separate analyses on the counselor and client side, we manually label the conversation turns as either counselor or client. Note that the speaker labeling can also be performed using automatic diarization, however since automatic speaker labeling on medical data is very challenging we decided to conduct this step manually.

[1] Also known as standardized patients.

Label	Words		Turns		Words/turn	
	Mean	Std	Mean	Std	Mean	Std
Counselor						
High	811	500	25	18	36	16
Low	423	368	15	13	33	19
Client						
High	674	439	25	18	31	17
Low	273	253	15	13	22	19
All	1168.4	837.0	42	34	31	14

Table 2: Word and turn statistics for low and high-quality sessions by counselors and clients in the counseling dataset.

Table 1 shows transcript excerpts of the two types of counseling conversations in the dataset. Word statistics of the final transcription set are provided in Table 2.

3.3 Annotation of Counseling Skills

In addition to our empirical assessments of counseling quality, for each conversation in the final set, we obtain standard measurements of MI adherence, which can be used as a proxy of MI quality. Specifically, we annotate two micro-skills that are frequently used in the evaluation of MI counseling: reflective listening and questions (Tollison et al., 2008). During this process, we used the Motivational Interviewing Treatment Integrity (MITI) coding scheme (Moyers et al., 2016), which is the current gold standard for MI fidelity evaluation.

The annotation was conducted by two undergraduate students who were trained in the use of the MITI 4.0. Before annotating the full set, we measured the reliability of the coding on a sample of 20 double-coded conversations, with an even

Code	Sample utterance
Question	What do you think it would take to change your mind about participating in physical activity?
Reflection	It sounds like you're concerned by your weight and you want to start to make positive changes.

Table 3: Verbal examples of questions and reflections in the dataset

distribution for the low-quality and high-quality categories. The resulting intra-class correlation scores for both Questions and Reflections are 0.96 and 0.94, respectively, thus showing good levels of agreement between the two annotators. Next, we split the remaining sessions among the two annotators to be coded independently. During the coding process, the annotators used both the audio recording and the transcript. The annotation was conducted at conversation turn-level using RQDA, an R package for Qualitative Data Analysis.[2]

The final annotation set consists of 1,981 questions and 1,180 reflections. Sample reflections and questions in our dataset are shown in Table 3.

4 Conversational Analyses

The goal of this paper is not only to learn models able to predict counseling quality but also to gain a better understanding of what makes a good counselor. Since the quality of counseling interventions relies highly on the active collaboration between counselors and clients, we analyze several linguistic aspects of the conversation in relation to counseling quality. Specifically, we focus on language exchange patterns over the conversation, sentiment trends, linguistic alignment, and topics discussed during the conversation.

4.1 Interaction at Turn Level

To explore the differences between low-quality and high-quality counseling we start by analyzing the counselor and client dialog interaction. Specifically, we analyze their turn-by-turn interaction by examining the average number of words used by each speaker as well as their word ratio. To visualize these aspects over time, we divide the session into five stages of nearly identical number of turns. Then, we calculate the average number of words per turn up to each stage.

The motivation for this split is to treat the counseling process as a sequence of continuous up-to-now information. Figures 1 and 2 show the words per turn by counselors and clients respectively.[3]

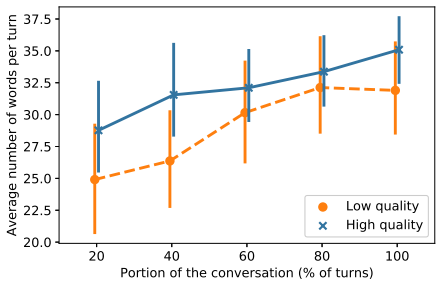

Figure 1: Average words per turn by counselors as the conversation progresses.

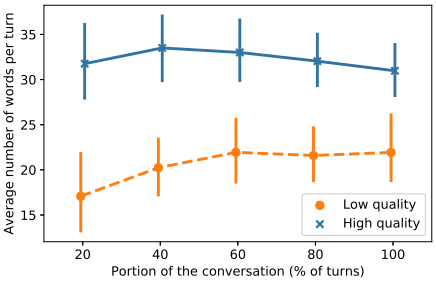

Figure 2: Average words per turn by clients as the conversation progresses.

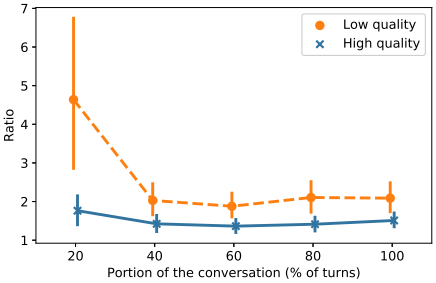

Figure 3: Word ratio by turn between counselors and clients as the conversation progresses.

From these graphs, we observe significant differences in the way clients participate in the counseling dialog, with clients speaking substantially less in low-quality conversations. Conversely, counselors seem to speak more during low-quality conversations; however, this difference is not statistically significant as there is a noticeable overlap between the two plots ($p < 0.05$; bootstrap resampling test). To take a closer look at the word exchange trends during the conversation, we also plot the counselor to client word ratio per conversation turn. The graph, shown in Figure 3, not only confirms this result but also suggests a more balanced word exchange between counselors and clients during high-quality interactions.

[2] http://rqda.r-forge.r-project.org/

[3] The error bars shown in all graphs are calculated using bootstrapping with a 95% confidence interval.

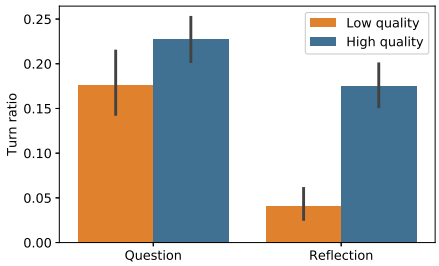

Figure 4: Number of questions and reflections per turn during low quality and high quality counseling.

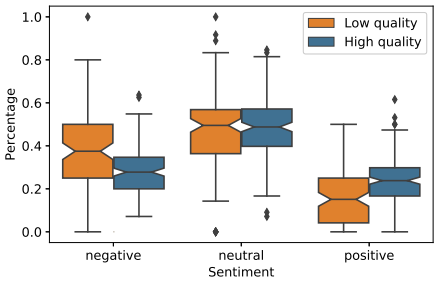

Figure 5: Distribution of positive, negative, and neutral sentiment by counseling quality.

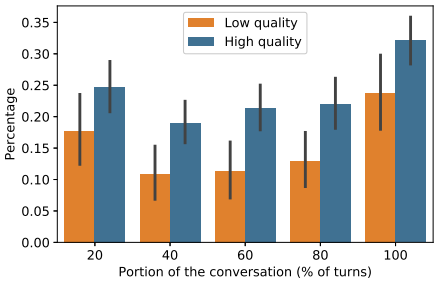

Figure 6: Distribution of positive sentiment over 20% splits of the conversation.

In addition, we examine whether counselor skills, measured by the questions and reflections asked during the conversation, differ based on the quality of the counseling. We thus compare the count of questions and reflections in the two types of counseling normalized by the number of turns in the session. The purpose of normalization is to filter out the effect of length and focus on the density of the behaviors. Figure 4 shows the results. The plot shows that reflection density in low-quality counseling is significantly lower than in high-quality counseling. On the other hand, question density does not seem to be significantly different between the two groups. These results are in line with previous findings of reflective listening being a skill associated with high-quality counseling (Glynn and Moyers, 2010)

4.2 Sentiment Trends

The sentiment expressed by counselors during the conversation can provide important insights into whether counselors focus on positive or negative aspects of client communication. We thus analyze the sentiment expressed across the conversation in relation to conversation quality. Given the effort required to manually annotate the sentiment in each conversation turn, we opt for using an automatic off-the-shelf sentiment classifier from the Stanford Core NLP package (Manning et al., 2014). Using this tool, we obtain the sentiment score for each conversation turn. The score ranges from very negative to very positive $--$, $-$, 0, $+$ and $++$, representing five sentiment categories in the order of increasing positiveness. Since $--$ and $++$ rarely occur in our dataset we treat both $--$ and $-$ as negative, 0 as neutral, and $+$ and $++$ as positive. Figure 5 shows the distribution of the three sentiment categories in low- and high-quality counseling respectively, where we observe that neutral sentiment occurs most frequently while the positive sentiment occurs the least.

We also examine sentiment changes throughout the conversation. As before, we divide the turns in each session into five splits and calculated the frequency of positive, negative and neutral sentiment at each stage. Figure 6 shows the positive sentiment trend in both low-quality and high-quality conversations.[4] The graph shows that during high-quality counseling, the counselors tend to express more positive sentiment than counselors in low-quality counseling. This suggests that positive language is a particular strategy of good counselors as they seem to show higher levels of positive sentiment across the conversation as compared to counselors in low-quality encounters. Furthermore, the U-shaped curve observed for the distribution of positive sentiment over the course of the conversations also points to counselors being more positive and friendly at the beginning of the conversation and ending the conversation with positive remarks.

5 Linguistic Alignment

The degree of language coordination that speakers show during a conversation is an indicator of whether they are able to establish a successful interaction (Pickering and Garrod, 2004). We examine the counselor and client language coordination

[4]We did plot the negative and neutral sentiment, however, we did not find significant differences between the two groups.

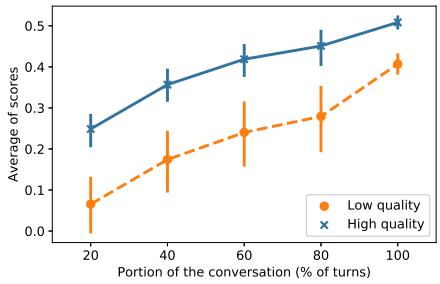

Figure 7: Linguistic style matching across five equal segments of the conversation duration.

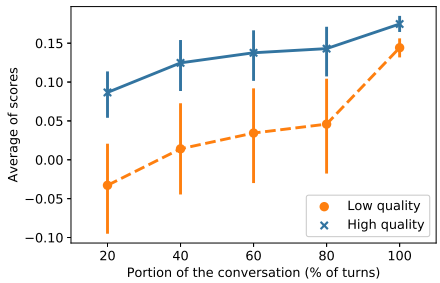

Figure 8: Linguistic style coordination across five equal segments of the conversation duration.

using the Linguistic Style Matching (LSM) (Gonzales et al., 2009) and Linguistic Style Coordination (LSC) metrics. These metrics quantify to which extent one speaker, i.e., the counselor, matches the language of the other, i.e., the client. Both metrics are evaluated at turn-level across eight linguistic markers from the LIWC dictionary (Tausczik and Pennebaker, 2010), including quantifiers, conjunctions, adverbs, auxiliary verbs, prepositions, articles, personal pronouns, and impersonal pronouns. Figures 7 and 8 show each metric trends across the two types of counseling.

Overall, the graphs suggest that during high-quality interactions, counselors show higher levels of linguistic alignment. This trend is more noticeable at the turn by turn level where steady and increased levels of linguistic matching are observed. Interestingly, this analysis shows the opposite behavior from the study by (Althoff et al., 2016), where successful counselors acted more in control over the conversation. We attribute this difference to two important aspects: 1) the type of conversation being analyzed, i.e., synchronous (face-to-face) vs asynchronous communication (text messages); and 2) the counseling strategy being implemented, i.e., in our case MI counseling, which is client-centered and thus the counselor's role is more supportive than directive.

6 Topics Discussed During the Conversations

During behavioral counseling, counselors should make an active effort to have a good understanding of the client's values and priorities (Miller, 1995). We explore how the conversation content relates to counseling quality by examining the topics discussed by counselors and clients. For this task, we apply the Meaning Extraction Method (MEM) (Chung and Pennebaker, 2008), a topic extraction method that identifies the most common words used in a set of documents, and cluster them into coherent themes by analyzing their co-occurrences. This method has been used in the past in the psychotherapy domain to analyze salient topics in depression forums (Ramirez-Esparza et al., 2008) and also to investigate differences in topics discussed by clients given their therapy outcomes (Wolf et al., 2010).

We first identify unique words that are exclusively used by either counselors or clients. During this process, we applied part of speech tagging to the counselor and client speech and remove domain related nouns such as drinking, alcohol, and smoking in order to obtain general topics. We also remove words with a frequency lower than five and keep words that appear at least in 5% of the speaker turns. Using the resulting word lists, we generate counselor and client matrices containing binary vectors indicating the use of each word by a specific speaker. Finally, we run a Principal Component Analysis (PCA), followed by varimax rotation on each document matrix to find clusters of co-occurring nouns. This process results in 10 and 8 components (topics) for counselors and clients respectively, explaining at least 35% of the total variance. To identify which topics are more salient for each speaker we use the method proposed in (Wilson et al., 2016) to measure the degree to which a particular MEM topic (component) is used during high-quality and low-quality encounters.

Results are shown in Table 4, which shows the scores assigned to each topic. In this table, scores greater than 1 correspond to topics salient in high-quality counseling, while scores lower than 1 indicate topics salient in low-quality counseling.

The analyses hint on interesting differences in the topics used by counselors during the conversations. In general, during high-quality interactions the counselors seem to focus on the client's mo-

Counselor			Client		
Component	Sample nouns	Score	Component	Sample nouns	Score
Motivation	Future, motivate, list, scale	1.62	Plan	Follow, focus,decide, plan, mind	1.17
Plan	Happen, explain, share, strategy, manage, successfully	1.27	Change	Difficult, lifestyle, kid, busy, ready, manage, important, realize	1.15
Importance	Importance, aware, scare, relate	1.25	Experience	Experience, past, happen, situation	1.06
Encourage	Confidence, experience, change, follow, support	1.22	Social	Parent, college, friend, group, school	1.04
Reasons	Concern, appreciate, tough, bring, interest	1.21	Concerns	Scary, risk, trouble, concern, upset, worse	1.02
Reflection	Connect, difficult, different, significant, tough, sound, deal	1.19	Family	Husband, child, mom, son, daughter, kid, house, love	0.99
Social	Parent, family, social, friend, child, role, wife, people	1.16	Reasons to Change	Habit, interest, break, consider, hard, important, sure	0.94
Persuade	Wrong, avoid, consequence, absolutely, worse	0.93	Uncertainty	Choice, confuse, young, deal	0.80
Plan	Plan, discuss, commit, focus	0.77			
Time	Talk, start, today, time, work	0.68			

Table 4: Themes used by clients and counselors during counseling conversations, along with sample nouns and salient theme scores. In this table, scores greater than 1 correspond to topics salient in high-quality counseling conversations while scores lower than 1 indicate topics salient in low-quality conversations.

tivations, reasons to change, and encouragement. Similarly, the clients discuss topics reflecting their desire to change and describe their experiences and concerns. In contrast, low-quality counseling shows more persuasion and uncertainty. Finally, scores closer to 1 indicate that regardless of the counseling quality, both counselors and clients discuss social and family topics as potential drives for change, which further confirms the use of MI in the conversations.

7 Discriminating Between Low- and High-quality Counseling

7.1 Linguistic Features

We explore the use of linguistic cues to build a computational model that predicts the overall quality of 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, i.e., ngrams. The features are extracted from the conversation transcripts.

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 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 words belonging to semantic categories that are potential markers for conversation quality.

Metafeatures: We also extracted a set of metafeatures that describe the conversation interaction, including the number of counselor turns, the number of client turns, the 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. The set includes the percentage of positive, negative, and neutral turns, the number of times the sentiment changes during the conversation i.e., positive to negative, negative to positive, and positive/negative to neutral, as well as counts of sequences increasing and decreasing sentiment intensity.

Linguistic Alignment: We measure the LSM and LSC metrics as described in section 4 over 74 LIWC categories and measured at 20% increments of the encounter duration.

Discourse topics: These features consist of the 10 topics identified in section 4 as frequently discussed during the MI encounters by counselors. The features are obtained by using regression-based factor scores.

MITI behaviors: This set includes the number of

Feature set	Counseling Quality		
		F-score	
	Acc.	Low	High
Baseline	59.846%		
N-grams	87.259%	0.849	0.890
Semantic	80.309%	0.763	0.832
Metafeatures	72.587%	0.297	0.830
Sentiment	74.517%	0.298	0.844
Alignment	72.587%	0.640	0.779
Topics	81.081%	0.768	0.840
MITI Behav	79.537%	0.787	0.808
All features	**88.031%**	0.857	0.897

Table 5: Overall prediction results and F-scores for counseling quality using linguistic feature sets

reflections and questions by the counselor during the conversation as well as the ratio of reflections to questions. The counts were derived from the manual annotations described in section 3.

7.2 Classification Results

We use the different feature sets described in section 7.1 to build classifiers that distinguish between high-quality and low-quality counseling. The experiments are performed using Support Vector Machine classifiers and evaluated using leave-one-out cross-validation. Our choice of the classification algorithm is motivated by the relatively small size of our dataset, which makes neural-based approaches less effective as they rapidly overfit. As a reference value, we use a majority class baseline, obtained by selecting high-quality as the default class label, which corresponds to 59.84% accuracy.

Table 5 shows the classification performance obtained when using each feature set at the time. We measure the performance of the classifiers in terms of accuracy and F-score, which provide overall and class-specific performance assessments. Compared to the majority baseline, all the feature sets demonstrate a clear improvement in the classification of counseling quality. Among all feature sets, N-grams attain the best performance, followed by discourse topics and the semantic feature sets. Furthermore, the combination of all the features sets achieve the best accuracy values.

To better understand the contribution of the different feature sets to the overall classifier performance, we conduct an ablation study, where we remove one group of features at a time. We perform feature ablation only for the features sets that represent linguistic aspects identified as good discriminators of counseling quality during our exploratory analyses.[5] Table 6 shows the classification results obtained when removing the alignment features, the topics, the MITI behaviors, and the N-grams. The removal of the linguistic alignment features showed an important drop in accuracy values, thus confirming our findings from Section 4 that linguistic alignment plays an important role in counseling quality. Excluding topic features does not seem to affect the model, suggesting that these features might provide redundant information already captured by the other features. More importantly, the results show that the automatically generated features can provide comparable performance to manually coded features (MITI behaviors) as the model does not show performance loss when removing this set. Finally, we also experimented with removing the N-gram features, which lead to the highest drop in performance, hence showing the importance of these features in the model.

Feature set	Counseling Quality		
		F-score	
	Acc.	Low	High
All features	**88.031%**	0.857	0.897
– Alignment	86.100%	0.836	0.879
– Topics	88.031%	0.857	0.897
– MITI Behaviors	88.031%	0.857	0.897
– N-grams	76.448%	0.702	0.805

Table 6: Feature ablation study

8 Conclusions

We presented an extensive analysis of linguistic aspects of the collaboration process during counseling conversations in relation to counseling quality. We specifically analyzed participants' turn-by-turn interaction, linguistic alignment, and topics discussed, as well as the sentiment expressed by the counselor during the conversation. Our main findings are summarized below.

Turn-by-turn interaction: During high-quality counseling, counselors achieve a more balanced word exchange with clients as the conversation progresses. This was also confirmed by our analysis of counseling micro-skills, which showed that good counselors use more reflective listening, thus suggesting that they speak less and listen more. In contrast, during low-quality conversations counselors tend to speak more than their clients thus

[5]The sentiment trends (sentiment) and turn-level metrics (metafeatures) did show important differences between groups but were not as stable as the other cues.

making it difficult to understand their needs.

Sentiment: Good counselors tend to express more positive sentiment than less successful counselors, which suggest that they focus on the positive aspects of the conversations rather than on the negative aspects.

Linguistic alignment: Good counselors mirror the language of their clients as high-quality interactions showed higher levels of linguistic alignment. This trend is more noticeable at turn-by-turn level where steady and increased levels of linguistic mirroring were observed.

Topics: Good counselors discuss topics related to behavior change and commitment whereas their counterparts focus more on resistance and persuasion. However, the general trend is to discuss topics related to family and social interactions, regardless of the counseling quality.

The results of our analyses were used to build accurate counseling quality classifiers that rely on linguistic aspects, with accuracies of up to 88% as compared to a majority baseline of 60%. Our experimental results showed that the proposed features can provide comparable performance to manually coded features for this task, thus potentially bypassing the need for manual annotations, which are usually costly and time-consuming. This is an important finding as an open problem in the counseling field is the need for computational tools that allow scaling-up the evaluation of the quality of MI interventions (Atkins et al., 2014).

In the future, we plan to build upon the acquired knowledge and the developed classifiers to generate systems able to provide actionable feedback on how to achieve high-quality counseling. Such systems can aid the process of acquiring or improving counseling skills for both novice and experienced MI counselors.

Finally, an important contribution of this work is the dataset collected, which is publicly available at `http://lit.eecs.umich.edu/downloads.html`.

Acknowledgements

This material is based in part upon work supported by the Michigan Institute for Data Science, by the National Science Foundation (grant #1815291), and by the John Templeton Foundation (grant #61156). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the Michigan Institute for Data Science, the National Science Foundation, or John Templeton Foundation.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics*.

Timothy R Apodaca, Brian Borsari, Kristina M Jackson, Molly Magill, Richard Longabaugh, Nadine R Mastroleo, and Nancy P Barnett. 2014. Sustain talk predicts poorer outcomes among mandated college student drinkers receiving a brief motivational intervention. *Psychology of Addictive Behaviors*, 28(3):631.

David C Atkins, Timothy N Rubin, Mark Steyvers, Michelle A Doeden, Brian R Baucom, and Andrew Christensen. 2012. Topic models: A novel method for modeling couple and family text data. *Journal of family psychology*, 26(5):816.

David C Atkins, Mark Steyvers, Zac E Imel, and Padhraic Smyth. 2014. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, 9(1):49.

Dogan Can, Panayiotis G. Georgiou, David C. Atkins, and Shrikanth S. Narayanan. 2012. A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features. In *INTERSPEECH*, pages 2254–2257. ISCA.

Cindy K Chung and James W Pennebaker. 2008. Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of Research in Personality*, 42(1):96–132.

Jacques Gaume, Gerhard Gmel, Mohamed Faouzi, and Jean-Bernard Daeppen. 2009. Counselor skill influences outcomes of brief motivational interventions. *Journal of Substance Abuse Treatment*, 37(2):151 – 159.

Lisa H Glynn and Theresa B Moyers. 2010. Chasing change talk: The clinician's role in evoking client language about change. *Journal of substance abuse treatment*, 39(1):65–70.

Amy L Gonzales, Jeffrey T Hancock, and James W Pennebaker. 2009. Language style matching as a predictor of social dynamics in small groups. *Communication Research*.

Edward Gottheil, Charles Thornton, and Stephen Weinstein. 2002. Effectiveness of high versus low structure individual counseling for substance abuse. *The American journal on addictions*, 11(4):279–290.

Zac E Imel, Scott A Baldwin, John S Baer, Bryan Hartzler, Chris Dunn, David B Rosengren, and David C Atkins. 2014. Evaluating therapist adherence in motivational interviewing by comparing performance with standardized and real patients. *Journal of consulting and clinical psychology*, 82(3):472.

Florian E Klonek, Vicenç Quera, and Simone Kauffeld. 2015. Coding interactions in motivational interviewing with computer-software: What are the advantages for process researchers? *Computers in Human Behavior*, 44:284–292.

Sarah Peregrine Lord, Elisa Sheng, Zac E Imel, John Baer, and David C Atkins. 2015. More than reflections: empathy in motivational interviewing includes language style synchrony between therapist and client. *Behavior therapy*, 46(3):296–303.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

William R Miller and Stephen Rollnick. 2013. *Motivational interviewing: Helping people change, Third edition*. The Guilford Press.

Theresa B Moyers, Tim Martin, Jon M Houck, Paulette J Christopher, and J Scott Tonigan. 2009. From in-session behaviors to drinking outcomes: a causal chain for motivational interviewing. *Journal of consulting and clinical psychology*, 77(6):1113.

Theresa B. Moyers, Lauren N. Rowell, Jennifer K. Manuel, Denise Ernst, and Jon M. Houck. 2016. The motivational interviewing treatment integrity code (miti 4): Rationale, preliminary reliability and validity. *Journal of Substance Abuse Treatment*, 65(Supplement C):36 – 42. Motivational Interviewing in Substance Use Treatment.

Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

Nairan Ramirez-Esparza, Cindy K. Chung, Ewa Kacewicz, and James W. Pennebaker. 2008. The psychology of word use in depression forums in english and in spanish: Testing two text analytic approaches. In *In Proc. ICWSM 2008*.

Stephen Rollnick, William R Miller, Christopher C Butler, and Mark S Aloia. 2008. Motivational interviewing in health care: helping patients change behavior. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 5(3):203–203.

C. Strapparava and A. Valitutti. 2004. WordNet-Affect: an affective extension of WordNet. In *Proceedings of LREC*, volume 4, pages 1083–1086.

Michael Tanana, Kevin Hallgren, Zac Imel, David Atkins, Padhraic Smyth, and Vivek Srikumar. 2015. Recursive Neural Networks for Coding Therapist and Patient Behavior in Motivational Interviewing. *2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 71–79.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Sean J. Tollison, Christine M. Lee, Clayton Neighbors, Teryl A. Neil, Nichole D. Olson, and Mary E. Larimer. 2008. Questions and reflections: The use of motivational interviewing microskills in a peer-led brief alcohol intervention for college students. *Behavior Therapy*, 39(2):183 – 194.

Amanda M Vader, Scott T Walters, Gangamma Chenenda Prabhu, Jon M Houck, and Craig A Field. 2010. The language of motivational interviewing and feedback: counselor language, client language, and client drinking outcomes. *Psychology of Addictive Behaviors*, 24(2):190.

Steven R. Wilson, Rada Mihalcea, Ryan L. Boyd, and James W. Pennebaker. 2016. *Cultural influences on the measurement of personal values through words*, volume SS-16-01 - 07, pages 314–317. AI Access Foundation.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, HLT-Demo '05, pages 34–35. Association for Computational Linguistics.

Markus Wolf, Cindy K Chung, and Hans Kordy. 2010. Inpatient treatment to online aftercare: e-mailing themes as a function of therapeutic outcomes. *Psychotherapy Research*, 20(1):71–85.

Bo Xiao, Daniel Bone, Maarten Van Segbroeck, Zac E Imel, David C Atkins, Panayiotis G Georgiou, and Shrikanth S Narayanan. 2014. Modeling therapist empathy through prosody in drug addiction counseling. In *Fifteenth Annual Conference of the International Speech Communication Association*.