# Predicting Counselor Behaviors in Motivational Interviewing Encounters

**Verónica Pérez-Rosas[1], Rada Mihalcea[1], Kenneth Resnicow[2], Satinder Singh[1], Lawrence An[3], Kathy J. Goggin[4] and Delwyn Catley[5]**

[1]Computer Science and Engineering, University of Michigan
[2]School of Public Health. University of Michigan
[3]Center for Health Communications Research, University of Michigan
[4,5]Department of Pediatrics. Children's Mercy Kansas City, University of Missouri–Kansas City
[1,2,3]{vrncapr,mihalcea,kresnic,baveja,lcan}@umich
[4,5]{kjgoggin,dcatley}@cmh.edu

## Abstract

As the number of people receiving psychotherapeutic treatment increases, the automatic evaluation of counseling practice arises as an important challenge in the clinical domain. In this paper, we address the automatic evaluation of counseling performance by analyzing counselors' language during their interaction with clients. In particular, we present a model towards the automation of Motivational Interviewing (MI) coding, which is the current gold standard to evaluate MI counseling. First, we build a dataset of hand labeled MI encounters; second, we use text-based methods to extract and analyze linguistic patterns associated with counselor behaviors; and third, we develop an automatic system to predict these behaviors. We introduce a new set of features based on semantic information and syntactic patterns, and show that they lead to accuracy figures of up to 90%, which represent a significant improvement with respect to features used in the past.

## 1 Introduction

Effective behavioral counseling is an essential element in combating public health issues such as mental health, substance abuse, and nutrition among others. A key component in training addiction counselors and other health care providers is providing detailed clinical feedback and evaluation. In clinical psychotherapy, this is done through behavioral coding, a labor intensive and time consuming process that requires highly trained practitioners who observe the counseling interactions via audio/video or reading transcripts and, then provide detailed evaluative feedback based on a set of predefined behaviors.

Recently, research efforts have been made towards implementing automatic means to assist this process and provide clinicians with tools to code and analyze counseling narratives (Atkins et al., 2014; Xiao et al., 2014; Klonek et al., 2015). Such tools can enable analyzes at larger scale by providing faster, cheaper, and more reliable methods for coding and data summarizing tasks.

Following this line of work, this paper presents a text-based approach for the automatic coding of counselor's verbal behaviors during counseling encounters. We focus our analysis on counseling conducted using Motivational Interviewing (MI), a well established evidence-based psychotherapy style, and the Motivational Interviewing Treatment Integrity (MITI) coding scheme.

## 2 Background on Motivational Interviewing

Miller and Rollnick define MI as a collaborative, goal-oriented style of psychotherapy with particular attention to the language of change (Miller and Rollnick, 2013). MI has been widely used as treatment method in clinical trials on psychotherapy research to address addictive behaviors such as alcohol, tobacco and drug use; promote healthier habits such as nutrition and fitness; and help clients with psychological problems such as depression and anxiety (Rollnick et al., 2008; Pollak et al., 2010; Lundahl et al., 2010; Vader et al., 2010; Apodaca et al., 2014; Magill et al., 2014; Moyers and Martin, 2006; Moyers et al., 2009; Glynn and Moyers, 2010; Barnett et al., 2014). In addition, MI has been successfully applied in different practice settings including social work in behavioral health centers, education, and criminal justice (Wahab, 2005; McMurran, 2009).

MI implementation requires effective counselor



*Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1128–1137,
Valencia, Spain, April 3-7, 2017. 

training, supervision, and evaluation. Counselor's competence in MI delivery is measured by either focusing on counselor behaviors, client behaviors, or both (Jelsma et al., 2015).

The MITI coding system is currently the gold standard instrument for this task (Moyers et al., 2005). MITI focuses on counselors verbal behaviors and measures their MI proficiency by evaluating the use of reflective listening; questions; counselor strategies to engage clients such as seeking collaboration, affirming, and emphasizing autonomy; behaviors that indicate counselor deficiencies while delivering MI such as confronting and persuading without permission; and finally, neutral behaviors such as providing information and persuading with permission.

## 3 Related Work

Recently there have been a number of efforts on building computational tools that assist clinical psychotherapy on behavioral coding tasks.

(Can et al., 2012) proposed a linguistic based approach to automatically detect and code counselor reflections that is based on analyzing n-grams patterns, similarity features between counselor and client speech, and contextual meta-features, which aim to represent the dialog sequence between the client and counselor. A method based on labeled topic models is presented in (Atkins et al., 2012; Atkins et al., 2014), where authors focus on automatically identifying topics related to MI behaviors from the MISC scheme such as reflections, questions, support, and empathy. Unlike their work, we introduce and experiment with richer sets of features that represent more accurately the linguistic structure of counselor behaviors, including syntactic patterns and semantic information. Moreover, although we also focus on the recognition of the two most frequently encountered behaviors (reflections and questions), we also apply and evaluate our system on the other MI behaviors measures by the MITI coding scheme. Speech and linguistic based methods have also been proposed to evaluate overall MI quality. For instance, (Xiao et al., 2014) presents a study on the automatic evaluation of counselor empathy. The method is based on analyzing correlations between prosody patterns and empathy showed by the therapist during the counseling interactions.

Although most of the work on coding of MI within session language has focused on modeling the counselor language, there is also work that investigates the client language. (Tanana et al., 2015) addresses the identification of counselor's statements discussing client's change talk. Their approach uses recursive neural networks to model sequences of counselor and client verbal exchanges. (Lord et al., 2015b) analyze the language style synchrony between therapist and client during MI encounters. They rely on the psycholinguistic categories from the Linguistic Inquiry and Word Count lexicon to measure the degree in which counselor language matches the client language.

Also related to our research is work on the social interaction domain. (Danescu-Niculescu-Mizil et al., 2012) studied power differences from language coordination in group discussions by measuring the similarity of word usage across different linguistic categories. Stylistic influence and symmetry have also been explored in social media interactions (Danescu-Niculescu-Mizil et al., 2011). More recently, (Althoff et al., 2016) explored these phenomena in the mental health domain by analyzing text-message-based counseling and observed that counselors who are more successful act with more control in the conversations and coordinate in a lower degree than their less successful counterparts.

In summary, research findings have shown that natural language processing approaches can be successfully applied to clinical narratives for the automatic annotation and analysis of therapists' and clients' behaviors. However, developed methods have not yet explored the use of linguistic features that incorporate semantic or syntactic information. In this paper we seek to explore new linguistic representations that can improve the identification of MITI counselor behaviors. Furthermore, we also experiment with features that measure participants linguistic accommodation during the counseling interaction.

## 4 MI Narratives Dataset

The data used in this study consists of 277 MI sessions conducted in several medical settings, including randomized control trials in clinical research for smoking cessation and medication adherence; MI training from a graduate-level MI course; wellness coaching phone calls; and brief medical encounters in dental practice and student

counseling. The full set comprises 97.8 hours of audio with an average session length of 20.8 minutes with a standard deviation of 11.5 minutes.

### 4.1 Transcription

Before transcribing, all the counseling recordings were preprocessed to remove any personal identifiers. This includes manually trimming the audio to remove introductions and replacing references to participant's name and location with silences.

Sessions were transcribed via Mechanical Turk (Marge et al., 2010) using the following guidelines: 1) transcribe speech turn by turn, 2) clearly identify the speaker (either client or counselor), 3) include speech disfluences, such as false starts, repetitions of whole words or parts of words, prolongations of sounds, fillers, long pauses. Transcriptions were manually verified at random points to avoid spam and ensure their quality. The final transcript set contains 22,719 utterances.

### 4.2 MI Coding

MITI coding was conducted by a team of three counselors who have extensive experience with MI.[1] Prior to the annotation phase, annotators participated in a coding calibration step where they discussed the criteria for sentence parsing, the correct assignment of behavior codes, and conducted team coding in a set of sample sessions.

As suggested in the MI literature, we evaluated the coding reliability on a sample of ten double-coded sessions, which were coded by our staff and by MITI developers (Moyers et al., 2005).

We measured the inter-annotator agreement at both session and utterance level. For the session level, we measured the Intraclass Correlation Coefficient (ICC), which indicates how much of the total variation in MITI scores is due to differences among annotators (Dunn et al., 2015). The utterance level agreement was measured using the Kappa score (Lord et al., 2015a).

The ICC values reported in Table 1 show noticeable high agreement for the Question and Reflection codes with scores ranging between 0.89 to 0.97, which are considered excellent agreement in the MI literature (Jelsma et al., 2015). The remaining codes show lower agreement values due to low frequency counts in the sample. This was particularly the case for the Giving Information, Affirm and Emphasizing Autonomy codes, for which we were unable to obtain ICC scores (NA). Confront, and Persuading without Permission codes are not reported as they did not appear in our sample. The main reason for this is that the dataset was derived from sessions conducted by experienced counselors who avoided such codes as they indicate bad MI practice.

Overall, the ICC scores suggest that the annotators do not show significant variations at session level coding, i.e., the total frequency counts of each code per session did not differ significantly between coders. Furthermore, the Kappa scores suggest that annotators have fair to good pairwise agreement at utterance level coding.

Since the inter-reliability analysis showed reasonable agreement among the coding team members, we moved forward to the annotation phase. The 277 sessions are randomly distributed among the three members of the coding team. Annotations are conducted using the session audio recording along with its transcript using Nvivo,[2] a quantitative analysis suite for behavioral coding that allows selecting free text and assigning it to a given category. Table 2 presents an excerpt of a session transcript. As observed, a talk-turn can comprise multiple utterances.

The team annotated approximately 20 sessions per week. The entire annotation process took nearly three months. After the annotation phase, the annotated transcripts were processed to extract the verbal content of each MITI annotation; non-coded utterances were also extracted and labeled as neutral speech. In the coded set 33% (5262) are Questions, 17% (2690) are Simple Reflections, 18% (2876) are Complex Reflections, and 32% (5058) are other MITI codes: Seeking Collaboration (614), Emphasizing Autonomy (141), Affirm (499), Confront (141), Persuading without Permission (598), Giving information (1017), and Persuading with Permission (2100).

## 5 Linguistic Features for MI behaviors

In order to explore linguistic patterns related to counselor behaviors, we analyze their definitions and usage. For instance, the use of reflective statements helps counselors understand client's statements through hypothesis testing (Miller and Rollnick, 2013); questions help counselor elicit information and engage the clients in the conversation;

[1] Annotators were trained in the use of MITI 4.1 by expert trainers from the Motivational Interviewing Network of Trainers

[2] http://www.qsrinternational.com/what-is-nvivo

| | Transcript | Code |
|---|---|---|
| T | *So, before we go further, you know, there's two different aspects of, you know, of weight. So there's the food aspect and the exercise aspect. Is there something that you'd particularly like to focus on today?* | GI,QUEST |
| C | Well I think my-my biggest concern is the food issue, and h-how to eat better. So, I think I'd like to start there. | |
| T | Okay. *Because you mentioned that you've been active before in sports* | S-REFL |
| C | Yeah, and, you know, with the nice weather coming, I'd like to get outside and do things, so I'm sure that will come, you know, soon. | |
| T | Right. | |
| C | I just, you know, it's so hard to-to change my eating habits. | |
| T | *So, it sounds like, you know, you may even feel, sort of, more confident that you'll be more active physically. And then that's why you'd like to focus on the food part. Because if you know that that's coming up and you're sure that you will be able to do that, then the food part would really help.* | C-REFL |
| C | Yeah, exactly. | |

Table 2: Transcript excerpt from a MI session between therapist (T) and client (C). MI codes include: Complex Reflection (C-Refl), Simple Reflection (S-Refl), Question (Quest), Giving information (GI). Coded utterances are shown in italics.

| Behavior | Inter-reliability | |
|---|---|---|
| | ICC | Kappa |
| Question | 0.97 | 0.64 |
| Complex reflection | 0.97 | 0.49 |
| Simple reflection | 0.89 | 0.34 |
| Seeking collaboration | 0.03 | 0.42 |
| Giving Information | NA | 0.28 |
| Affirm | NA | 0.47 |
| Emphasizing autonomy | NA | 0.31 |

Table 1: Inter-annotator agreement for the MI dataset in a random sample of 10 sessions

and so on. Considering these guidelines, we derive the following features that aim to capture the linguistic differences among these behaviors.

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

*Semantic information*: These features attempt to bring semantic information into the analysis of counselor language by identifying words as belonging to certain semantic categories that are potential markers of counseling style. For instance, semantic categories related to reflective language include tentative language e.g., maybe, perhaps, looks, as well as anxiety words e.g., afraid, tense, worried. We use two groups of semantic features. The first consists of features derived from the LIWC lexicon (Tausczik and Pennebaker, 2010), a psycholinguistic resource that contains 70 semantic categories representing psychological cues to human thought processes, emotional states, intentions, and motivations. The second is a self-acquired reflection lexicon consisting of 146 words frequently present during reflective statements. These features are represented as the total frequency counts of all the words in a word category that are present in the annotation.

*Similarity*: Since reflective listening includes repetition and rephrasing, we can expect to observe linguistic similarity between client and counselor speech. Thus, we measure the degree to which the counselor matches the client language by using Linguistic Style Matching (LSM) (Gonzales et al., 2009), a technique that allows to quantify the extent to which one person uses comparable types of words to another person. We measure LSM at a turn-by-turn level using the LIWC word categories, e.g., positive words, pronouns, negations, quantifiers. In order to capture information from return statements, we combine client speech from the previous and current turn along with the counselor utterance. The features are represented by a score ranging between 0 and 1 indicating the degree to which the counselor and client use the same type of words.

*Syntactic features*: These features aim to represent the syntactic structure of the clinician statements. We use these features to encode information about the word order in the sentence. We expect syntactic patterns with high occurrence will likely capture reflection starters commonly used

| Features | Acc. | P | R | F |
|---|---|---|---|---|
| REFL vs. ALL | | | | |
| Baseline | 75.00% | 0.00 | 0.00 | 0.00 |
| N-grams | 82.47% | 0.69 | 0.66 | 0.67 |
| Semantic | 77.37% | 0.68 | 0.34 | 0.45 |
| Similarity | 60.95% | 0.58 | 0.34 | 0.42 |
| Syntax | 78.65% | 0.72 | 0.67 | 0.62 |
| All features | 82.62% | 0.69 | 0.66 | 0.67 |
| REFL vs. OTHER | | | | |
| Baseline | 64.00% | 0.00 | 0.00 | 0.00 |
| N-grams | 83.51% | 0.78 | 0.80 | 0.79 |
| Semantic | 71.57% | 0.68 | 0.54 | 0.58 |
| Similarity | 63.00% | 0.58 | 0.27 | 0.37 |
| Syntax | 86.80% | 0.82 | 0.86 | 0.84 |
| All features | 81.27% | 0.84 | 0.83 | 0.84 |

Table 3: Classification results for counselor reflections (REFL), other MITI codes (OTHER), other MITI codes + transition (unannotated) utterances (ALL)

by the counselor such as "it sounds like ...". First, we use the Stanford parser to generate the Context Free Grammar parse trees of counselor utterances and extract all production rules present in the trees. Second, we derive features for each lexicalized and unlexicalized production rule augmented with its grandparent node; this means we also include chunk tags such as noun phrases, adverb phrases, prepositional phrases, and so on. Third, each feature is calculated by counting how many times a production rule or production-rule-sequence occurs in the utterance.

## 6 Experiments and Results

After the feature extraction, we explore whether these features can be used as predictors of counselor behaviors. We first focus on the prediction of reflections and questions, as they represent the most frequent behaviors in counseling narratives; we then experiment with the use of these features for the prediction of the other behavior codes.

### 6.1 Predicting Counselor Reflections

We conduct learning experiments where we explore the use of n-grams, syntactic, semantic, and similarity features to build reflection classifiers at three levels of detail.

First, we attempt to mimic the process human coders follow while MITI annotating a session, i.e., the coder goes through each counselor utterance and chooses the most appropriate code according to the MITI guidelines. Hence, we focus on the identification of Reflection utterances regardless of being complex or simple. The learning task aims to classify a counselor utterance either as a Reflection, or a Not-reflection, i.e., any other counselor utterance. Second, given that a large portion of the verbal exchanges between the counselor and the client consists of transition or facilitative statements (e.g., yeah, right), we decided to remove this content from the analysis thus focusing on the task of discriminating between Reflection and any other MITI code. Third, we aim to discriminate between Simple and Complex reflection. In MI, counselors use both types of reflections to understand the client's perspective, feelings, and values. However, in general, complex reflections are preferable over simple reflections as they show counselor's deeper understanding of the issues being discussed. In a real setting, distinguishing between these two behaviors and understanding their linguistic differences is important in order to provide the counselor with feedback on the nature of their reflective statements.

During our experiments we employ the Support Vector Machines (SVM) (Cortes and Vapnik, 1995) classifier as the main classifier. We use the version implemented in the LibLinear library with the default parameters. We build several classification models using each of the different sets of linguistic features. We evaluate the ability of such models to predict the target behavior using a five-fold cross-validation. As reference value, we use a majority class baseline, which is the percentage of instances correctly classified when selecting by default the most frequent category in the training data.

Table 3 summarizes the classification performance for each set of features in the detection of reflections. During our experiments we used F-score as the main evaluation metric. This metric considers both the proportion of reflections identified from the training set (recall) and the proportion of reflections correctly identified as such (precision). From this table, we can observe that the syntactic model accurately captures differences between reflective and non-reflective content. This difference is even more noticeable when discriminating between syntactic structures associated to reflective statements versus syntactic structures associated to other MITI codes (REFL vs OTHER column).

Table 4 presents the classification performance for the simple (S-Refl) and complex reflections (C-Refl). F-scores among the classification mod-

| Features | Acc. | S-REFL | | | C-REFL | | |
|---|---|---|---|---|---|---|---|
| | | P | R | F | P | R | F |
| Baseline | 52.00% | 0.00 | 0.00 | 0.00 | 0.52 | 1.00 | 0.68 |
| N-grams | 63.24% | 0.61 | 0.65 | 0.64 | 0.66 | 0.62 | 0.63 |
| Semantic | 67.21% | 0.63 | 0.65 | 0.64 | 0.67 | 0.70 | 0.69 |
| Similarity | 63.22% | 0.62 | 0.59 | 0.58 | 0.67 | 0.58 | 0.63 |
| Syntax | 65.06% | 0.63 | 0.66 | 0.64 | 0.67 | 0.70 | 0.657 |
| All features | 62.52% | 0.60 | 0.62 | 0.61 | 0.64 | 0.62 | 0.63 |

Table 4: Classification results for Simple Reflections (S-Refl) and Complex Reflections (C-Refl)

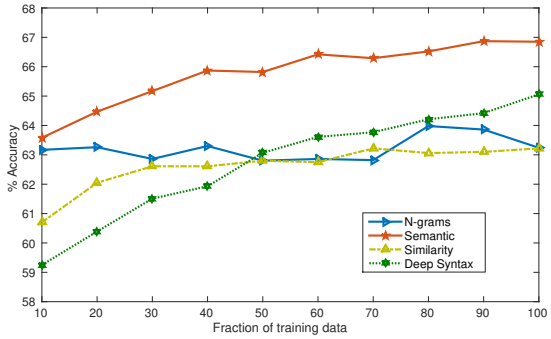


Figure 1: Learning curve for Simple (S-Refl) and Complex Reflection (C-Refl) detection using different amounts of training data and four feature sets.

els do not show noticeable improvement gain for the syntactic model. This suggest that the syntactic features might not have enough discriminative power to accurately differentiate between Complex and Simple reflections as the differences among these behavior codes are mostly semantic. This could be also attributed to the similarity of syntactic constructions for Simple and Complex reflections, i.e., common starters such as *it sounds like*, *it looks like*, and other similar syntactic constructions. Note that this task is also challenging for humans, as reported pairwise inter-annotator agreements for Simple and Complex reflections in our sample ranges between fair to good levels (see Kappa values in Table 1).

Finally, we investigate whether larger amounts of training data can be helpful to discriminate between Simple and Complex reflections, in particular with the use of syntax-based features. We plot the learning curves of the different sets of features using incremental amounts of data as shown in Figure 1. The learning trend suggests the classification performance while distinguishing between Complex and Simple Reflection improves when increasing the number of training examples. Notably, the syntactic features curve shows consistent growth, suggesting that larger quantities of training data might improve the classification performance for this task.

| Target Behavior Change | Sessions | REF | OTHER |
|---|---|---|---|
| Medication adherence | 93 | 2977 | 4031 |
| Smoking Cessation | 95 | 2290 | 3745 |
| Dietary Changes | 72 | 2045 | 2669 |

Table 5: Class distribution for three target behavior changes

### 6.2 The Role of Behavior Change Target

During MI encounters, counselors follow specific strategies to guide the client towards behavior change. For instance, reflective listening strategies include generic starters to reformulate, rephrase, or intensify client's statements, which are used regardless of the behavior change target. Considering that MI has proven to be effective on addressing a wide range of application domains, this might suggest a certain degree of domain independence, which can be of importance for the development of natural language processing strategies for the automatic coding of MI sessions.

Aiming to explore the role played by the health issue being discussed during the counseling encounter, we conduct an additional set of experiments on three target behavior changes present in our dataset, namely medication adherence, smoking cessation, and dietary changes. The class distribution for each set is shown in Table 5. We exclude 16 sessions as they correspond to miscellaneous change goals.

Using this data, we build reflection classifiers using the linguistic features described before. Results are shown in Table 6. From this table, we can derive interesting observations. First, following a similar trend as in our previous experiments, syntactic features offer improved performance over the n-grams, similarity, and semantic feature sets. Second, we observe more steady improvement when using a combination of different feature sets, as compared to our previous experi-

| Target Behavior Change | Baseline | N-grams | Semantic | Similarity | Syntax | All Features |
|---|---|---|---|---|---|---|
| Medication adherence | 57.52% | 83.56% | 61.48% | 57.52% | 85.31% | 88.78% |
| Smoking cessation | 62.05% | 82.41% | 66.16% | 62.96% | 83.41% | 85.34% |
| Dietary changes | 56.51% | 82.75% | 66.50% | 59.56% | 82.42% | 85.41% |

Table 6: Classification performance for reflection detection (REFL VS. ALL) on sessions aiming at three behavior changes

| Category | Medication Adherence | Smoking Cessation | Dietary Changes |
|---|---|---|---|
| Noun | health, family, routine, life | children, control, past, role, parent | pre-diabetes, people, husband, future |
| Verb | live, imagine, see, eat, help | see, affect, feel, want, talk | care, affect, concern, leave |
| Adjective | difficult, responsible, important | concern, hard, helpful | scary, busy, difficult, willing |

Table 7: Counselor word usage across different health issues

ments. Still, the performance for both sets of experiments is comparable, thus suggesting that the task is not heavily affected by the health issue being discussed. This further suggests that training data on the same behavior target is desirable but not essential.

Since we did not observe noticeable differences in language constructions used by counselors across different health issues, we decided to analyze whether counselors differ in their word choices. We thus looked at the top syntactic features generated for each classification model and their corresponding terminal nodes and part of speech tags. Table 7 shows sample words for nouns, verbs, and adjectives used by counselors while formulating reflections for three target behavior changes. From this table we notice that counselor word usage does vary with the health issue being addressed. For instance, when discussing smoking cessation, counselor emphasize verbs and nouns that evoke clients' desire to change (affect, want, feel) and discuss client values that are related to how they are perceived by others (role, parent, control).

### 6.3 Prediction of Counselor Questions

Our next set of experiments aims to predict counselor questioning statements. As before, we build different prediction models using the developed feature sets and attempt to discriminate between 1) Questions and any other counselor utterance (QUEST vs ALL), and 2) Questions and other MITI codes (QUEST vs OTHER). Classification performances for these models are shown in Table 8. The best performing feature set is the syntactic followed by the n-grams model.

Note that in addition to the experiments reported in this table, we attempted to combine different features sets. However, we did not observe substantial improvement over our best performing model consisting of syntactic features.

| Features | Acc. | P | R | F |
|---|---|---|---|---|
| QUEST vs. ALL | | | | |
| Baseline | 76.83% | 0.00 | 0.00 | 0.00 |
| N-grams | 87.81% | 0.91 | 0.92 | 0.76 |
| Semantic | 79.19% | 0.69 | 0.37 | 0.48 |
| Similarity | 62.33% | 0.36 | 0.31 | 0.36 |
| Syntax | 90.59% | 0.82 | 0.81 | 0.81 |
| All features | 81.87% | 0.76 | 0.74 | 0.75 |
| QUEST vs. OTHER | | | | |
| Baseline | 66.87% | 0.00 | 0.00 | 0.00 |
| N-grams | 88.84% | 0.86 | 0.85 | 0.85 |
| Semantic | 75.28% | 0.70 | 0.64 | 0.67 |
| Similarity | 57.33% | 0.38 | 0.49 | 0.42 |
| Syntax | 90.48% | 0.92 | 0.92 | 0.87 |
| All features | 86.57% | 0.82 | 0.82 | 0.82 |
| GRU | 92.8% | 0.89 | 0.92 | 0.90 |

Table 8: Classification results for counselor questions (QUEST), other MITI codes + transition (unannotated) utterances (ALL), and other MITI codes (OTHER).

### 6.4 Prediction of Other MI Codes

Aiming to identify potential predictors for the remaining MI codes, we conduct a set of experiments where we use our linguistic feature sets to build multiclass classifiers. Table 9 shows the precision and recall figures obtained for the different classification models. Note that the results using semantic and similarity features are not reported, as the resulting classifiers showed very low recall values. From these results we observe that both syntactic features and n-grams aid the identification of other counselor behavior codes, particularly Giving Information, Affirm, and Seeking Collaboration. However, the prediction accuracy of the syntactic models is slightly lower than the n-grams models. We believe that the more verbose nature of these codes, in contrast to reflections, makes it difficult to benefit from syntactic patterns.

| | GI | | AF | | SEEK | | AUTO | | PWP | | PWOP | | CON | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Features | P | R | P | R | P | R | P | R | P | R | P | R | P | R |
| N-grams | 0.56 | 0.47 | 0.42 | 0.37 | 0.55 | 0.49 | 0.29 | 0.26 | 0.23 | 0.15 | 0.29 | 0.22 | 0.10 | 0.07 |
| Syntax | 0.52 | 0.47 | 0.41 | 0.35 | 0.54 | 0.50 | 0.26 | 0.19 | 0.18 | 0.14 | 0.28 | 0.22 | 0.10 | 0.04 |
| GRU | 0.48 | 0.79 | 0.28 | 0.75 | 0.41 | 0.80 | 0.24 | 0.30 | 0.09 | 0.47 | 0.21 | 0.38 | 0.05 | 0.17 |

Table 9: Classification results for seven MI behaviors: Giving Information (GI), Affirm (AF), Seeking Collaboration (SEEK), Emphasizing Autonomy (AUTO), Persuading with Permission (PWP), Persuading without Permission (PWOP), Confront (CON)

| | QUEST | REFL | S-REFL | C-REFL |
|---|---|---|---|---|
| 1 | *. ^ S → ? | *VBZ ^ VP → sounds | ROOT ^ ROOT → S | VP ^ S → TO_VP |
| 2 | *. ^ SBARQ → ? | *IN ^ SBAR → since | *VBD ^ VP → mentioned | NP ^ PP → PRP$ _NN |
| 3 | ROOT ^ ROOT → SBARQ | *IN ^ S → so | ROOT ^ ROOT → FRAG | *VBZ ^ VP → sounds |
| 4 | NP ^ SQ → PRP | S ^ ROOT → IN _NP | ROOT ^ ROOT → NP | *VB ^ VP → be |
| 5 | *.SQ → ? | VP ^ S → VBZ _SBAR | VP ^ S → VBD _SBAR | *TO ^ VP → to |
| 6 | ROOT ^ ROOT → SQ | *PRP ^ NP → it | *RB ^ ADVP → so | *RB ^ ADVP → really |
| 7 | *WP ^ WHNP → what | *RB ^ ADVP → really | VP ^ S → VBD _NP | *PRP ^ NP → it |
| 8 | *IN ^ PP → about | S ^ ROOT → CC _ADVP | *NN ^ NP → ok | *IN ^ SBAR → like |
| 9 | *DT ^ NP → any | ADJP ^ VP → RBR _JJ | *UH ^ INTJ → okay | *IN ^ PP → like |
| 10 | *. ^ FRAG → ? | VP ^ S → VBP _PRT | VP ^ S → VBD _PP | *DT ^ NP → this |

Table 10: Most discriminative syntactic features for Questions (QUEST), Reflections (REFL), Simple reflection (S-REFL) and Complex reflection (C-REFL).

Also, note that Emphasizing Autonomy, Persuading With And Without Permission, and the Confront codes lead to low precision and recall values, which can be partially attributed to having a smaller number of training examples as compared to the other codes.

## 7 Discussion

Our experimental results support the use of automatic means to predict MITI counselor behaviors. Unsurprisingly, better results are obtained for the more frequent behaviors such as reflections and questions. Unlike previous studies that focused on the identification of reflective content in psychotherapy narratives (Atkins et al., 2014; Xiao et al., 2014), we build prediction models that predict all the MITI behavior codes. We also introduce and leverage new features consisting of semantic and syntactic patterns; our experimental results suggest the effectiveness of these new features.

To gain further insight into the syntactic patterns, we extract the most predictive features for each classification model. Table 10 presents a summary of the top ten production rules associated to Question (Quest), Reflection (Refl), Simple Reflection (S-Refl), and Complex Reflection (C-Refl). As expected, question production rules include the question mark as a clear indicator of questions. However, we also observe clause tags and phrase tags that capture more complex questioning structures such as direct questions introduced by a wh-word or a wh-phrase *SBARQ*, inverted yes/no questions *SQ*, question personal pronoun *WP* (wh-pronoun, personal), and noun phrases *WHNP*.

Similarly, the most predictive rules for Reflections (REFL column) include adverbs (*RB, RBR*), adjectives (*JJ*), present tense verbs (*VBZ*), personal pronouns (*PRP*); as well as adverb and adjective phrases (*ADVP, ADJP*). We observe that the syntactic similarity of reflective statements is well represented by the syntactic model as they include verbal structures that are frequently used by the counselor to formulate reflective statements; for instance, generic reflection starters such as "So, it sounds like ..." (see rules 1, 3, 5, and 6 in column REFL), and word categories, such as adjectives, conjunctions and comparative adverbs (see rules 8, 9, and 10 in column REFL). Moreover, we observe an interesting difference in the verb tense usage for Simple and Complex Reflection detection: production rules for Simple Reflection include present tense while production rules for Complex Reflection include past tense.

Overall, our experimental results show the potential of applying linguistic methods in the prediction of counselor behaviors, and in particular those that incorporate syntactic information into the analysis.

## 8 Conclusions

In this paper, we presented a classification model towards the automation of MI coding using the MITI coding system.

We made two important contributions. First, we introduced a novel large psychotherapy dataset derived from MI interventions, consisting of 277 MI sessions with a total of 22,719 utterances. The dataset was manually transcribed and annotated with ten counselor verbal behaviors. Second, using several features, we applied the classification model to the recognition of MI counseling behaviors, with an emphasis on the two most frequently encountered behaviors: reflections and questions. We showed how a richer feature set, and in particular a set consisting of semantic and syntactic patterns, can lead to accuracy figures of up to 90%, which represents a significant improvement with respect to the bag-of-word features used in the past.

We also presented several analyses, including an exploration of the role of the behavior change target in the prediction of reflections; and an analysis of the most discriminative features in the syntactic model. Although this study focused on the MITI as the coding system and MI as the counseling approach, we believe that the proposed methods could apply to other measures of MI skill fidelity such as Behavior Change Counselor Index (BECCI) (Lane et al., 2005), Independent Tape Rating Scale (ITRS) (Martino et al., 2009), Stimulated Client Interview Rating Scale (SCIRS) (Arthur, 1999), and the One Pass coding system (McMaster and Resnicow, 2015).

## Acknowledgments

This material is based in part upon work supported by the University of Michigan under the M-Cube program, by the National Science Foundation (grant #1344257), the John Templeton Foundation (grant #48503), and the Michigan Institute for Data Science. 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 University of Michigan, the National Science Foundation, the John Templeton Foundation, or the Michigan Institute for Data Science. We gratefully acknowledge the help of the three counselors who helped with the data annotations.

## 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 Arthur. 1999. Assessing nursing students' basic communication and interviewing skills: the development and testing of a rating scale. *Journal of Advanced Nursing*, 29(3):658–665.

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.

Elizabeth Barnett, Theresa B. Moyers, Steve Sussman, Caitlin Smith, Louise A. Rohrbach, Ping Sun, and Donna Spruijt-Metz. 2014. From counselor skill to decreased marijuana use: Does change talk matter? *Journal of substance abuse treatment*, 46(4):498–505.

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.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words! Linguistic style accommodation in social media. In *Proceedings of WWW*, pages 745–754.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of WWW*, pages 699–708.

Chris Dunn, Doyanne Darnell, Sheng Kung Michael Yi, Mark Steyvers, Kristin Bumgardner, Sarah Peregrine Lord, Zac Imel, and David C. Atkins. 2015. Should we trust our judgments about the proficiency of motivational interviewing counselors? a glimpse at the impact of low inter-rater reliability. *Motivational Interviewing: Training, Research, Implementation, Practice*, 1(3):38–41.

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*.

Judith G.M. Jelsma, Vera-Christina Mertens, Lisa Forsberg, and Lars Forsberg. 2015. How to measure motivational interviewing fidelity in randomized controlled trials: Practical recommendations. *Contemporary clinical trials*, 43:93–99.

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.

Claire Lane, Michelle Huws-Thomas, Kerenza Hood, Stephen Rollnick, Karen Edwards, and Michael Robling. 2005. Measuring adaptations of motivational interviewing: the development and validation of the behavior change counseling index (becci). *Patient education and counseling*, 56(2):166–173.

Sarah Peregrine Lord, Doğan Can, Michael Yi, Rebeca Marin, Christopher W. Dunn, Zac E. Imel, Panayiotis Georgiou, Shrikanth Narayanan, Mark Steyvers, and David C. Atkins. 2015a. Advancing methods for reliably assessing motivational interviewing fidelity using the motivational interviewing skills code. *Journal of substance abuse treatment*, 49:50–57.

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

Brad W. Lundahl, Chelsea Kunz, Cynthia Brownell, Derrik Tollefson, and Brian L. Burke. 2010. A meta-analysis of motivational interviewing: Twenty-five years of empirical studies. *Research on Social Work Practice*.

Molly Magill, Jacques Gaume, Timothy R. Apodaca, Justin Walthers, Nadine R. Mastroleo, Brian Borsari, and Richard Longabaugh. 2014. The technical hypothesis of motivational interviewing: A meta-analysis of mis key causal model. *Journal of consulting and clinical psychology*, 82(6):973.

Matthew Marge, Satanjeev Banerjee, Alexander Rudnicky, et al. 2010. Using the amazon mechanical turk for transcription of spoken language. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5270–5273. IEEE.

Steve Martino, Samuel A. Ball, Charla Nich, Tami L. Frankforter, and Kathleen M. Carroll. 2009. Informal discussions in substance abuse treatment sessions. *Journal of substance abuse treatment*, 36(4):366–375.

Fiona McMaster and Ken Resnicow. 2015. Validation of the one pass measure for motivational interviewing competence. *Patient education and counseling*, 98(4):499–505.

Mary McMurran. 2009. Motivational interviewing with offenders: A systematic review. *Legal and Criminological Psychology*, 14(1):83–100.

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

Theresa B. Moyers and Tim Martin. 2006. Therapist influence on client language during motivational interviewing sessions. *Journal of substance abuse treatment*, 30(3):245–251.

Theresa B. Moyers, Tim Martin, Jennifer K. Manuel, Stacey M.L. Hendrickson, and William R. Miller. 2005. Assessing competence in the use of motivational interviewing. *Journal of substance abuse treatment*, 28(1):19–26.

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.

Kathryn I. Pollak, Stewart C. Alexander, Cynthia J. Coffman, James A. Tulsky, Pauline Lyna, Rowena J. Dolor, Iguehi E. James, Rebecca J. Namenek Brouwer, Justin R.E. Manusov, and Truls Østbye. 2010. Physician communication techniques and weight loss in adults: Project chat. *American journal of preventive medicine*, 39(4):321–328.

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.

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. *NAACL HLT 2015*, page 71.

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.

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.

Stephanie Wahab. 2005. Motivational interviewing and social work practice. *Journal of Social Work*, 5(1):45–60.

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*.