Modeling Motivational Interviewing Strategies On An Online Peer-to-Peer Counseling Platform

RAJ SANJAY SHAH*, Georgia Institute of Technology, USA
FAYE HOLT*, Georgia Institute of Technology, USA
SHIRLEY ANUGRAH HAYATI, Georgia Institute of Technology, USA
AASTHA AGARWAL, Georgia Institute of Technology, USA
YI-CHIA WANG, Independent Researcher, USA
ROBERT E. KRAUT, Carnegie Mellon University, USA
DIYI YANG, Stanford University, USA

Millions of people participate in online peer-to-peer support sessions, yet there has been little prior research on systematic psychology-based evaluations of fine-grained peer-counselor behavior in relation to client satisfaction. This paper seeks to bridge this gap by mapping peer-counselor chat-messages to motivational interviewing (MI) techniques. We annotate 14,797 utterances from 734 chat conversations using 17 MI techniques and introduce four new interviewing codes such as "*chit-chat*" and "*inappropriate*" to account for the unique conversational patterns observed on online platforms. We automate the process of labeling peer-counselor responses to MI techniques by fine-tuning large domain-specific language models and then use these automated measures to investigate the behavior of the peer counselors via correlational studies. Specifically, we study the impact of MI techniques on the conversation ratings to investigate the techniques that predict clients' satisfaction with their counseling sessions. When counselors use techniques such as reflection and affirmation, clients are more satisfied. Examining volunteer counselors' change in usage of techniques suggest that counselors learn to use more introduction and open questions as they gain experience. This work provides a deeper understanding of the use of motivational interviewing techniques on peer-to-peer counselor platforms and sheds light on how to build better training programs for volunteer counselors on online platforms.

CCS Concepts: • **Human-centered computing** → **HCI theory, concepts and models**; *Empirical studies in collaborative and social computing*.

Additional Key Words and Phrases: Online Mental Health Communities, Natural Language Processing, Social Support

ACM Reference Format:
Raj Sanjay Shah, Faye Holt, Shirley Anugrah Hayati, Aastha Agarwal, Yi-Chia Wang, Robert E. Kraut, and Diyi Yang. 2022. Modeling Motivational Interviewing Strategies On An Online Peer-to-Peer Counseling Platform. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 527 (November 2022), 24 pages. https://doi.org/10.1145/3555640

*Both authors contributed equally to this research.

Authors' addresses: Raj Sanjay Shah, Georgia Institute of Technology, USA, rajsanjayshah@gatech.edu; Faye Holt, Georgia Institute of Technology, USA, mholt9@gatech.edu; Shirley Anugrah Hayati, Georgia Institute of Technology, USA, shirley@gatech.edu; Aastha Agarwal, Georgia Institute of Technology, USA, aagrawal319@gatech.edu; Yi-Chia Wang, Independent Researcher, USA, yichia.wang@gmail.com; Robert E. Kraut, Carnegie Mellon University, USA, robert.kraut@cmu.edu; Diyi Yang, Stanford University, USA, diyiy@stanford.edu.

2573-0142/2022/11-ART527
https://doi.org/10.1145/3555640

1 INTRODUCTION

Peer support occurs when non-professionals give and receive emotional support, shared knowledge, and shared social or practical help to others suffering from some problem [52]. Prior research indicates that peer support services can be an integral part of mental health recovery [48] and benefits support seekers [5], even though most peer counselors have not received professional training [13]. Online peer-to-peer support communities for mental health, such as Talklife [27], 7 Cups of Tea [40], and MellowTalk [33], have become more popular because they are easy to access and less costly than support from professional therapists [16]. The number of support seekers using online peer counseling platforms and online mental health communities (OMHCs) has grown significantly in the last few years [53] and has seen a surge in demand due to the COVID-19 pandemic [6]. To address the shift of many counseling service to peer-to-peer online formats, there is an increased need to explore how counseling can best adapt to online media. To fill this gap, this paper focuses on one such peer-to-peer counseling platform, using 7 Cups of Tea (7Cups) as a case study, to understand how peer-counselors behave, with a future goal of creating successful training programs for peer-counselors.

On 7Cups, volunteer counselors receive a short training session based on the client-centered guidelines of active listening and motivational interviewing [60]. Motivational interviewing (MI) is a set of client-centric counseling techniques used by therapists to implicitly direct patients to make lifestyle changes or perform actions towards their recovery [35]. MI focuses on collaboration between the volunteer counselors (known as listeners in 7Cups) and support seekers (known as members), in which the counselors interact with clients using specific skills (or techniques) [34–37]. Being able to identify and measure these skills at scale allows for a detailed analysis of strategies used in counseling sessions. Lundal et al. [29] have shown MI to be effective in bringing positive behavioral change for support seekers through a large-scale meta-analysis of 119 studies. Magill et al. [30] have shown the efficacy of MI-consistent techniques in promoting client language in support of behavior change. Conversely, MI-inconsistent techniques were shown to promote client language *against* behavior change, which is correlated with negative outcomes for clients [30].

We utilize client satisfaction as an outcome metric to explore the value of individual MI techniques, following prior work [1, 5, 25]. Although therapeutic alliance is a traditional metric used in many studies of the success of OMHCs [21, 31, 50], due to the complexity of measuring it [3], client satisfaction provides a useful proxy readily available as an outcome measure in many OMHCs. Additionally, client satisfaction is linked to therapeutic alliance, which is generally associated with retention and behavioral change [21, 31, 50]. On 7Cups, clients can choose a rating based on a 5-point Likert scale from 1 (worst) to 5 (excellent) to express their satisfaction of their conversations with listeners. In this work, we explore MI usage at a large-scale within an online medium. We are especially interested in MI techniques that predict client satisfaction and are correlated with client change talk from prior literature. The focus of our paper, therefore, is to understand which MI techniques that counselors use in OMHCs that lead to client satisfaction. Thus, we address the following two research questions:

RQ1: Which motivational interviewing techniques predict more satisfying conversations?
RQ2: How does listeners' use of motivational interviewing techniques change with their counseling experience?

To answer these questions, we have annotated a corpus of 734 conversations comprising 14,797 messages (utterances) with MI codes using a multi-label approach. Additionally, we defined new classes to capture other language behavior common in OMHCs. Manually labeling millions of utterances is a laborious task; therefore, we built machine learning classifiers to measure the MI and other language behaviors at scale. This paper uses 7Cups data to build mental-health

domain specific models. We then use the annotated corpus data to fine-tune these domain specific, transformer-based large language models [54] to predict the MI code of any given listener utterance. This allows us to perform our analysis at a scale of millions of conversations and utterances.

For RQ1, we first grouped conversations on 7Cups into two categories: "*satisfactory*" ones with rating of 4 or 5 and "*unsatisfactory*" ones with a rating of 1 to 3. Then we labeled each conversations' utterances using the fine-tuned classifiers and use logistic regression analysis to determine which MI techniques best predict whether conversations were satisfactory or unsatisfactory. RQ2, along with results from RQ1, tests whether listeners improve their behavior as they gain more experience. We analyzed this by identifying users who have been active for over a year. We then pinpoint which MI techniques become more/less prevalent as a listener gained experience.

While there is prior Human-Computer Interaction (HCI) research focusing on the identification of various strategies for online counseling success [1, 10], our work extends this by measuring both MI techniques frequently used offline and online-specific supportive strategies to predict conversational satisfaction. Results from millions of online conversations can extend the existing literature on the use of motivational interviewing by professional therapists to understand what works online.

2 RELATED WORK

2.1 Peer-to-Peer Counseling in Online Communities

Research has shown that peer support in OMHCs can lead to increased feelings of belonging, connectedness, and insight [38]. Additionally, peer-support can offer a solution to the historical inaccessibility of traditional mental health services, which has resulted in a majority of mental problems being untreated [23]. Prior research on peer-to-peer counseling has shown the efficacy of peer-counseling for clients. For example, recent work on 7Cups compared users' satisfaction with peer counseling to more formal psychotherapy and found that users' perceived peer-counseling to be as effective as psychotherapy [5]. Users also reported that they felt peer-counselors were more genuine than professional psychotherapists [5]. Graham et al. [19] has additionally shown the importance of peer-counseling for marginalized communities, such as the LGBTQ+ community and racial minorities.

The value of peer-counseling in OMHCs , however, probably depends upon the training and skill level of each peer-counselor. Under-trained volunteer counselors could lead to harmful sessions, as research has found that therapy strategies such as rigidity and over-control led to clients reporting they felt disempowered and devalued [11]. As such, there is a growing need for peer-counseling training programs on OMHCs. In offline settings, research conducted on peer-counselor training within a senior community reported a statistically significant improvement in peers' knowledge and skills following training [8]. While peer-counseling OMHCs vary in the degree and presence of listener training, some OHMCs such as 7Cups have introduced training programs for listeners based in active listening (an approach where counselors utilize open-ended questions to guide the conversation) and motivational interviewing [40, 60]. However, previous research on OMHC training programs found that volunteers on 7Cups reported feeling unprepared [60] and online volunteer listeners for OMHCs are typically not as well trained as volunteer listeners who conduct peer-counseling offline [44]. These discrepancies indicate the need for improved training programs that are tailored to the specific online environment where volunteer listeners will be working.

2.2 Identifying Techniques for Online Counseling Success

Due to the above described popularity of OMHCs, recent HCI research has focused on what counseling techniques lead to success [1, 10, 47, 59]. The counseling success of OMHCs has been

defined in terms of a variety of metrics, from user churn [56] and retention [59] to more complex measures of affective, behavioral, and cognitive psychosocial change [47]. Due to the complexity and nuance of measuring success within online counseling sessions, we focus on client satisfaction as our measure of success, in the tradition of prior work [1, 5, 25]. Furthermore, this self-reported metric is linked to the degree of therapeutic alliance, which is a standard metric in this area and associated with client retention [21, 31, 50]. Client satisfaction differs from therapeutic alliance, which specifically measures the relationship between support seeker and listener, whereas client satisfaction is only one of many potential indicators of therapeutic alliance. It is also important to note that even therapeutic alliance has certain shortcomings in measuring the overall "success" of therapeutic sessions; while therapeutic alliance is predictive of user engagement, retention, and behavior change during counseling, it is not clear whether therapeutic alliance is a predictor of clients sustaining behavioral changes after counseling [32].

Saha et al. [47] found the primary predictors of success are complex language factors such as adaptability and style. Similarly, consistent communication [59], linguistic style-matching [1], and concrete, positive, supportive feedback from listeners [10] are all reported indicators of successful counseling sessions. While past research focused on detecting patterns correlated to counseling success on OMHCs, not much attention has been paid to analyzing counseling sessions with a comprehensive framework, with a few exceptions like [4, 42, 43]. Our work extends this line to work by analyzing session satisfaction at scale through a predefined MI schema.

2.2.1 Client-Centered Counseling. Our research is based in a specific form of client-centered counseling that adds a directional component called motivational interviewing [35]. Client-centered counseling is a counseling approach developed by Carl Rogers that emphasizes active listening and a non-directive conversational style [7]. Client-centered strategies are effective in treating a variety of mental health problems such as addiction, anxiety, and depression [18] and are the current chosen approach for OMHCs such as 7Cups [40]. The quality of client-centered MI counseling sessions has been explored in offline counseling media through language modeling and fine-grained analyses of MI codes [18, 42, 43]. We extend from these fine-grained analyses to analyze how MI strategies apply to OMHCs.

Harrison and Wright conducted one such study on client-centered counselors within a text-based online environment and found counselors developed stylistic changes distinct from their in-person approaches, indicating there may be significant differences in client-centered strategies used online versus offline. Additionally, [45] conducted an analysis of client-centered counseling over video conferencing and reported counselors struggled more with establishing connections with clients and noted the increased difficulties for clients to perceive acceptance and empathy from their counselors. Despite these difficulties, [45] ultimately concluded that the condition for successful client-centered therapy can be met effectively in online media, pointing to the potential for client-centered therapy to be utilized effectively on peer-counseling platforms. To further explore the usage of client-centered counseling online and provides insights into how the challenges of online counseling may be effectively met through MI.

2.2.2 AI for Online Mental Health Communities. Researchers are starting to build integrated artificial intelligence (AI) application to improve the effectiveness of those participating in online support platforms [12, 41, 46]. Many technological aids to assist in online writings like MepsBot have been implemented in different domains [17, 55, 58] and have potential within OMHCs to aid volunteer counselors to improve sessions. Our research aims to offer insights that may potentially be incorporated into a similar AI writing assistant. While the direct incorporation of AI applications into OMHCs is still in preliminary stages, there have been several applications of AI to analyze linguistic patterns and explore factors that might be responsible for counseling success [1, 51].

AI has also been specifically applied to the domain of understanding and improving MI counseling offline. Atkins et al. [4] utilized topic modeling to predict MI codes. Imel et al. [20] explored the feasibility of an automated, machine-learning based feedback system for MI, and Flemotomos et al. [15] similarly explored automatically annotating MI codes to predict a therapist's performance both at the utterance and session levels. Our work continues this line further to explore text-based online therapy (versus dialogue) and create a new MI framework specific to online counseling. Our research builds on prior work focusing on automatic MI annotation and its applications and attempts to improve MI model performances and additionally answers several RQs that explore the nuances of online counseling.

3 MOTIVATIONAL INTERVIEWING TREATMENT INTEGRITY (MITI) FRAMEWORK

Our research is based on a client-centered counseling style called *motivational interviewing* [35]. Motivational interviewing has long been an effective treatment for substance abuse and addiction [35]. More recently, motivational interviewing has become a popular schema for peer-to-peer counseling in platforms such as 7Cups [40]. Given the effectiveness of motivational interviewing and its use in OMHCs, the goals of this paper are to measure MI skills at scale, assess which MI skills predict successful client evaluation of counseling sessions, and examine how use of these skills evolve over listeners' tenure. To do so, we use motivational interviewing skill codes, which are a set of predefined behavioral codes, to categorize listener utterances in a conversation [34]. To meet the unique requirements of an online peer-counseling platform, we create a MITI Framework of thirteen MI codes aggregated from three motivational interviewing manuals [34, 36, 37] and supplement them with four newly created codes discussed in Section 4.1. The MITI Framework is described and summarized in Table 1. The examples in the table are taken from conversations on 7Cups and paraphrased using round-trip translation to ensure anonymity.

Our framework follows the traditional MISC guidelines and divides the codes into three categories: (1) MI-consistent ones that promote the client-centered nature and goal of MI; (2) MI-inconsistent ones that undermine this goal either through inappropriate behavior or by explicitly telling the client what to do without collaborating or emphasizing the client's autonomy; and (3) Others that serve primarily to establish rapport with the client during a session [2]. In addition to categories defined in Table 1, we initially also included 4 other codes detailed in MI manuals - *Structure*, and the MI-inconsistent codes: *Warn*, *Confront* and *Raise Concern*. However, these classes rarely occurred in the 7Cups conversations we analyzed, with fewer than 50 per class among 14,797 listener utterances. With so few cases, they could not be modeled in the classification task and were therefore removed from our framework.

4 DATASET

7 Cups of Tea is an online platform that provides free counseling to those experiencing emotional distress [40] by connecting volunteer counselors to support seekers on an anonymous chatroom for counseling. The listeners on 7Cups are volunteers who receive 30-60 minutes of training and must pass an exam, but are not licensed professionals. This structure allows 7Cups to offer free, accessible counseling, but with minimally trained volunteers who may not be as equipped as professionals to handle members dealing with emotional distress and complex mental health problems. After each conversation, members have the option to rate their conversation with the listener on a 5-point Likert scale from 1 (worst) to 5 (excellent) to indicate how satisfied they were with the conversation.

This work obtains the chat messages sampled from the 7Cups platform for the annotation of the dataset, via a research collaboration with 7 Cups of Tea. The annotated dataset consists of 14,797 labeled unique listener utterances (or 734 conversations) from the platform. These listener utterances are labeled according to the motivational interviewing framework described in Table 1.

Code	Description	Examples
MI-Consistent		
Affirm	The listener says something positive or complimentary. It may serve as reinforcement.	*This is such a big step forward! I am very proud of you.*
Emphasizing Autonomy	The listener focuses responsibility on the member.	*OK, this is your call.*
Open Question	Questions that are not closed and leave a latitude for response.	*What brings you here today?*
Closed Question	A question that implies a short answer.	*Have you sought professional help with these triggers?*
Persuade (with Permission)	Listener makes overt attempt to change member's opinions, attitudes, or behavior. Listener asks for permission first or emphasizes collaboration.	*Unfortunately, I am not a doctor, but it is always best to get a second opinion from someone who is at least believable.*
Reflection	Reflections capture and return to the member something the member has said.	*You feel dysphoric and it causes you suffering.*
Seeking Collaboration	Listener explicitly attempts to share power or acknowledge the expertise of the member.	*May I ask what decisions do you feel bad about?*
MI-Inconsistent		
Direct	Counselor gives an order or command.	*You need to go somewhere else and get a second opinion.*
Inappropriate*	Excessive swear words, asking unnecessary personal information, abusive behavior, etc.	*What are you wearing?*
Other		
Grounding	Facilitate or acknowledge.	*Okay or Sure or Hmm*
Giving Information	Educate, provide feedback, or expresses a professional opinion without persuading, advising, or warning.	*There are some great meditation links on here.*
Support*	Sympathetic, compassionate, or understanding comments	*I'm always here to help you.*
Personal Disclosure	When the listener shares something personal as part of the conversation.	*I know exactly what you mean, I was like that a few years back.*
Introduction	Greetings, exchanging names, etc.	*Hello welcome to 7COT. My name's John*
Conclusion	Statement indicating that the listener need to leave.	*I'm ending our chat now.*
Chit-Chat*	Connecting with members by conversing about topics not related to the general session.	*I wonder if we can fly one day.*
Other	Any utterance that does not fit into the above.	

Table 1. MITI Framework; *denotes newly created codes.

The distribution of the MI techniques and their sizes are given in Table 2. Since we utilized a multi-label approach, in which each utterance may receive up to three MI code labels, the distribution counts in Table 2 sum to more than the 14,797 unique utterances labeled.

4.1 Motivational Interviewing Categories

Listener utterances are annotated using 17 class categories as presented in Table 2. Note that the distribution of classes is highly imbalanced. Some of the classes like *Open Questions*, *Closed Questions* and *Personal Disclosure* have more than 1500 datapoints each, whereas classes like *Emphasizing Autonomy*, *Inappropriate*, and *Direct* have less than 300 data points. This class imbalance is characteristic of client-therapist conversations and previous works have shown a similar trend [26]. This class skewness presents a challenge for developing listener behavior in natural language processing models. We use the one versus all binary classification approach described in Section 5 to deal with this class skewness during automated labeling. In addition to the motivational interviewing codes obtained from the MITI Manual [36], we have created 4 new codes to more

Category	#Utterances	Annotation Agreement (Krippendorff's alpha)
Giving Information	304	1
Reflection	1697	0.585
Affirm	346	0.659
Closed Question	1956	0.868
Open Question	2507	0.950
Persuade	1918	0.424
Direct	250	0.666
Emphasizing Autonomy	120	0.665
Grounding	1027	0.826
Personal Disclosure	1605	0.658
Introduction/Greeting	1260	0.877
Conclusion	340	0.791
Seeking Collaboration/Permission	271	0.662
Support	1493	0.842
Inappropriate	224	1
Chit-Chat	963	0.59
Other	1299	0.664

Table 2. Distribution of categories

accurately describe the listener utterances in this online context: *Support*, *Inappropriate*, *Chit-Chat*, and *Other*, and define them as follows:

- **Support:** This class describes utterances that are generally sympathetic, compassionate, or understanding comments. Utterances following the *Support* technique are characterized by "*agreeing with the client*". Few examples of *Support* are: "*You've got a point there*" (agreement), "*That must have been difficult*" (compassion), "*I can see why you feel this way*" (understanding).
- **Inappropriate:** This class refers to utterances that include swear words, asking unnecessary personal identifiable information, or abusive behavior. The creation of this class was especially motivated by 7Cups being an online, anonymous chat platform, which often includes inappropriate comments. Therefore, identifying these comments is an important step in creating a better environment for both clients and listeners.
- **Chit-Chat:** This class was also motivated by the unique medium of an online chat. *Chit-Chat* is a non-traditional therapeutic code that describes utterances where jokes are made or the listener otherwise attempts to connect to the client in a manner that is not directly related to mental health or the topic at hand. Although this strategy is not traditionally therapeutic, *Chit-Chat* may serve to make the client more comfortable with the listener and could potentially have positive affects on the conversation's success, which is a topic we intend to explore further.
- **Other:** This class serves as a label to capture utterances that do not fit accurately into the other 16 categories. Example utterances may be typos or short responses such as, "Good" that are not neutral enough to be labeled *Grounding*.

4.2 The Annotation Process

Building the corpus required several rounds of annotation. The initial annotation process had three components. First, three annotators individually labeled a series of conversations. Second, we calculated Krippendorff's alpha [24] for each label to measure agreement between annotators. We considered a Krippendorff's alphas of approximately 0.7 as the cutoff of a reasonable level of agreement for a specific MI label, building on the rule of thumb of the Krippendorff's alpha [24] score. Lastly, the annotators went through disagreements, clarified misconceptions, and decided on the correct label(s) for an utterance where there had been disagreement.

Once the Krippendorff's alphas reached the threshold of 0.7 for most of the classes (Table 2), we determined that there was a consistent, accurate understanding of the MI codes and there was enough agreement between the annotators to begin labeling separately. This allowed us to significantly speed-up the process of labeling utterances. Additionally, to further increase the rate of labeling, we supplemented manual labeling with a semi-supervised approach using machine learning classifiers based on a large-pretrained language model called BERT [14]. BERT is a neural network architecture that outperforms other architectures on a variety of downstream language processing tasks, such as classification, due to its ability to capture context specific semantics of language. BERT and BERT-like models capture task agnostic semantics of words due to the unsupervised pre-training on large amounts of textual information like Wikipedia articles. To speed up the annotation task, we further fine-tuned the BERT model to capture mental health specific word usage and the labeling process. For each batch of conversations, the listener utterances were inputted into each BERT classifier. We defined a threshold of 0.7 confidence and only looked at the utterances the classifiers predicted as belonging to a certain class with >= 0.7 confidence. This cutoff was chosen based on the annotator's qualitative analysis of the classifiers' outputted labels; it was observed that if a model labeled an utterance with <0.7 confidence, it was rarely, if ever, correct when manually verified. The annotators then manually verified each of the classifiers' outputs to ensure their correctness before assigning a label to the utterance. We found that this semi-supervised process was very effective for the labels: *Personal Disclosure*, *Grounding*, *Introduction*, *Closed Question*, and *Open Question*, most likely because these class rely less heavily on the context of previous chat messages than others. The remaining techniques were primarily labeled manually. This combination of manual and semi-supervised annotation allowed us to increase the corpus from 1,539 labeled listener utterances (or 147 conversations) to 14,797 labeled listener utterances (or 734 conversations).

5 AUTOMATED LABELING USING LARGE LANGUAGE MODELS

We implement automated labeling using large scale language models fine-tuned on the data described in Section 4. This allows us to analyze our research questions at the scale of millions of utterances. Constructing this automation pipeline required several considerations, described in the following paragraphs, to meet the nuances of an OHMC.

Language models learn language patterns from the data using statistical methods. Such models are useful for many machine learning prediction tasks. Recently, large scale language models trained on unlabeled textual data have achieved impressive results on classification tasks [14, 22, 28]. In this study, we train a classifier using these large language models and fine-tune it using our annotated dataset. This allows us to automatically label listener utterances with MI codes.

We use several state–of-the-art architectures of large-scale language models from the Huggingface library [57]: BERT-base-uncased (base-line), BERT-large-uncased [14], RoBERTa-base [28], Vinai/BERTweet [39]. Our experiments compare the models (Appendix Tables 7, 8) and show that Vinai/BERTweet architecture has the best F1 score as it is pre-trained on Twitter data which matches

Model Name	No previous context	1 previous utterance as context	5 previous utterances as context
BERT-Base-Uncased	0.570	0.600	0.539
BERT-large	0.605	0.639	0.608
RoBERTa-Base	0.608	0.628	0.606
BERTweet	0.642	0.693	0.649
Mental Health BERT	**0.678**	***0.725***	**0.705**

Table 3. Averaged F1 Scores across MI Codes for Different Models without Domain Specific Pretraining

MI Code	No context	1 previous utterance	5 previous utterances	Most Relevant Words
Giving Information	**0.656**	0.574	0.387	mean, tell, ok, aww, oh
Reflection	0.631	0.692	**0.780**	like, yeah, know, think, feel
Support	0.721	0.750	**0.778**	understand, okay, hope, feel, sorry
Affirm	**0.677**	0.624	0.392	good, great, nice, awesome, love
Closed Question	0.905	**0.915**	0.895	ok, talk, like, know, want
Persuade	0.661	0.770	**0.841**	like, think, know, happen, thing
Open Question	0.918	**0.943**	0.938	like, happen, tell, mean, think
Seeking Collaboration	**0.518**	0.500	NA	ask, tell, maybe, mind, want
Inappropriate	0.566	0.667	**0.783**	kiss, big, eat, wet, wanna
Direct	0.516	**0.523**	0.514	need, dont, stop, talk, tell
Emphasizing Autonomy	0.730	**0.785**	NA	choice, want, best, decide, right
Grounding	0.773	**0.846**	0.835	ok, yeah, oh, cool, hmmm
Personal Disclosure	0.803	0.776	**0.817**	good, know, want, yeah, time
Introduction	0.899	**0.927**	0.851	hello, good, morning, today, welcome
Conclusion	**0.806**	0.756	0.500	sleep, bye, care, gonna, later
Chit-Chat	0.307	**0.652**	0.497	NA
Other	0.458	0.633	**0.770**	NA
Averaged F1	0.678	**0.725**	0.705	NA

Table 4. Best Model Metrics (MH-BERT): F1 scores of the Positive Classes for MI Codes using 0, 1, and 5 utterances as previous context when using *domain specific* Mental Health BERT

the informal short text messaging format of the 7Cups platform. To capture domain specificity, we build a domain specific instance (known as Mental Health BERT) by further pre-training the Vinai/BERTweet checkpoint (best performing model) on 120 million chat messages from 7Cups.

For training the classifier, we build a classifier for each MI code using a one versus all method to deal with the large number of classes. Thus, an utterance can have multiple MI codes to address the issue of low representation of classes in our dataset. An example utterance for multiple MI codes would be: "*Hi, how are you doing today?*", which has two MI codes, *Introduction* and *Open Question*. For each of the 17 MI codes, we measure the F1 score of the positive class (the presence of an MI technique in an utterance) for evaluation. The metrics of accuracy, precision, and recall are not best suited for highly imbalanced data, especially for a binary classification task such as ours since models tend to overfit towards the class with greater representation.

Human annotation for some MI codes, such as *Support*, *Reflection*, and *Giving Information*, required previous context to determine the correct micro-skill for an utterance. To examine the

role of context in developing accurate machine learning models, we fine-tune them by adding the previous context in our data by looking at 0 (no previous context), 1, and 5 previous utterances. The addition of context is done by concatenating the previous k utterances to the current utterance (where k belongs to {0, 1, 5}). The context is independent of the person who wrote the utterance (i.e., the member or the listener). Using this setup, Table 3 gives the averaged F1 scores of the respective MI codes as positive classes. The F1 scores of MI codes as positive classes for the best model are shown in Table 4. The F1 scores of MI codes for other models are given in the Appendix (Table 7, 8). We highlight the key takeaways of the model development process:

- Best Architecture - Mental-Health-BERT (based on vinai/bertweet [39], Table 4): The MH-BERT model works substantially better than the models not trained on the domain specific data, as observed in Table 3.
- Optimum Previous Messages as Context: Table 3 shows that using 1 previous utterance for additional context has the best classification performance, regardless of the choice of model architecture.
- Inappropriate: For *Inappropriate*, we finetune HateBERT [9] for few shot classification, along with other models, as HateBERT is trained on abusive language that is fairly similar to the inappropriate text found in the 7Cups data. HateBERT gives an F1 score for the positive class of 0.6061 (No context), 0.7119 (1 previous utterance as context), 0.8421 (5 previous utterances as context). These metrics are much better than the other models discussed above.
- Presence of NA Values: Some of the cells in Table 4 contain NA values because there is not enough representation of the utterances in the annotated data for the particular MI technique to have 5 previous utterances. For example, *Seeking Collaboration* has only 27 positive data points when considering 5 previous utterances as context; this makes it very difficult to build robust classifiers due to limited training data.

As we use automated labeling without manual verification to explore our research questions, we seek to ensure that the quality of the labeling is maintained when automated. Therefore we randomly sampled 100 annotated utterances and manually analyzed the results. Two authors annotate the hundred samples, we calculate inter-rater agreement and the agreement with the model for validation. After comparing model labeling with human analysis, we find a Krippendorff's Alpha agreement score of 0.614 (Table 9). We observe perfect inter-rater agreement for certain MI codes due to their low representation in the data sample. As a sanity check, we also retrieve the top words of each MI code according to the term frequency-inverse document frequency scores. For this step, we removed stop words and lemmatized all the words for standardization of verb forms. The top words for each MI code are given in Table 4. The words generated by the natural language processing models are consistent with the definitions of the codes (Table 1), giving distinct but relevant top words for at least 10 out of the 15 codes under consideration (excluding *Chit-Chat* and *Other*). This is true for codes like *Introduction*, *Conclusion*, *Emphasizing Autonomy*, *Affirm*, *Direct*, *Open Questions*, *Closed Questions*, *Support*, *Inappropriate* and *Grounding*. We do not show words for *Chit-Chat* and *Other*, as these two MI techniques are meant to encapsulate utterances that are diverse in nature in terms of the topics discussed. These robustness checks confirm the efficacy of our automated pipeline.

6 ANALYSES

For our analyses, we utilize the Mental Health BERT model described above to predict MI codes on a sampled set of conversations from 7Cups taken over a six-month time-span (January 2020 to August 2020). Since we are interested in understanding what counseling strategies in a conversation leads to client satisfaction, we convert satisfaction ratings to two categories: *satisfactory* conversations

(received a rating or 4 or 5) and *unsatisfactory* conversations (received a rating of 1-3). The threshold of 4 for a satisfactory conversation was chosen based on the average rating across all conversations in our dataset being 4.14.

6.1 RQ1: What Motivational Interviewing Techniques Predict More Satisfaction?

To answer RQ1, we study the association between listener MI usage and satisfying conversations. Conversations are categorized into satisfactory or unsatisfactory ones, as displayed in Table 5. The average rating for these conversations is 4.14, indicating that members may be more likely to leave a rating if the conversation was satisfactory. In total, we analyze 63,012 conversations (5,632,402 utterances) to explore what MI techniques lead to conversation success.

6.1.1 Logistic Regression Analysis of MI Codes and Satisfactory Conversations. After obtaining MI label(s) for each utterance in each conversation, we run a logistic regression model to predict the outcome of a conversation (1 = satisfactory conversation, 0 = unsatisfactory conversation). We use the number of times each MI code occurs in a given conversation as our independent variables. To account for the subjectivity of using a rating scale from 1-5 (as certain members may be more likely to rate all conversations lower or higher), we use member's past average rating as a control variable. We also use listener age and member age as control variables and find in Table 6 that every control variable used is statistically significant, with member's average past rating having an especially high odds ratio of 1.685, indicating the importance of controlling for a member's rating tendency. To account for the imbalanced counts of our sample data listed in Table 5, the classes in our logistic regression model are weighted accordingly, using the inverse of the respective class frequencies.

Conversation Category	Count	Percent of Total Count
Satisfactory	49,892	79%
Unsatisfactory	13,120	21%

Table 5. Breakdown of samples used per conversation category for analysis in Section 6.1.

Table 6 shows the coefficients and odds ratio of each motivational interviewing technique resulting from the logistic regression model. We observe that *Reflection*, *Persuade*, and *Affirm* are statistically significant predictors for client satisfaction while *Inappropriate* is a statistically significant negative predictor for client satisfaction. *Reflection* is the most statistically significant positive predictor and has an odds ratio of 1.03, which indicates that each additional usage of *Reflection* in a conversation (a one unit increase) increases the odds of the conversations receiving a satisfactory rating by 3%. It is important to note that each of the significant positive predictors for client satisfaction (*Reflection*, *Persuade*, *Affirm*) are MI-consistent, aligning with previous research on the success of MI-consistent codes in standard offline counseling sessions [30].

In prior work, success within MI counseling is measured by the prevalence of "change talk" defined by Miller and Rolnick as, "any self-expressed language that is an argument for change" and a lack of "sustain talk" defined as, "the person's own arguments for not changing, for sustaining the status quo" [35]. A previous study on which MI codes were most likely to be followed by "change talk" and/or "sustain talk" found that *Reflection* was significantly more likely to be followed by client "change talk", aligning with our finding that *Reflection* is a significant indicator of client satisfaction [2]. However, this same research found that "sustain talk" was also more likely to follow usage of *Reflection*, which may indicate that while *Reflection* leads to client satisfaction, there are issues with this MI strategy in regards to its relationship with "sustain talk". The dual nature of

MI Code	Coefficient	Odds Ratio
MI-Consistent		
Affirm	0.036*	1.037
Emphasizing Autonomy	0.078 ⊙	1.081
Open Question	0	1
Closed Question	0.011 ⊙	1.012
Persuade (with permission)	0.018**	1.019
Reflection	**0.035*** **	**1.035**
Seeking Collaboration	0	0.999
MI-Inconsistent		
Direct	0.019	1.019
Inappropriate	**-0.086*** **	**0.917**
Other		
Grounding	0.008 ⊙	1.008
Giving Information	-0.025	0.975
Support	0.013	1.013
Personal Disclosure	0.001	1.001
Introduction	0	1
Conclusion	0.014	1.015
Chit-Chat	-0.004	0.996
Control Variables		
Member Age	-0.007***	0.993
Listener Age	-0.003***	0.997
Member Past Average Rating	0.522***	1.685
AIC of Model	97,800	

Table 6. Associations between strategies and satisfactory conversations. ***p < 0.001; **p<0.01; *p<0.05 ⊙p<0.1.

Reflection positively being followed by "change talk" and negatively being followed by "sustain talk" may be understood by some examples from 7Cups. The following examples have been paraphrased using round-trip translation to ensure anonymity: *'Life is hard' (Listener 1)* versus *'Everyone is different. Some make friends quickly, some don't. It will take time to adjust. (Listener 2)*. From these examples of *Reflection*, there is evidence that *Reflection* varies in usage. *Listener 1* will most likely contribute more to "sustain talk" as they are reflecting the member's current state of mind and not using *Reflection* as a reframing tool. *Listener 2*, conversely, reflects back the member's situation and reframes it in a more positive context, which is more likely to lead to "change talk". Therefore, we see from Table 6 and these qualitative examples that our findings on *Reflection* as a significant predictor of client satisfaction aligns well with prior research reporting *Reflection* correlated with "change talk" and "sustain talk".

The same prior work by Apodaca et al. [2] found that *Affirm* was the only MI code predicted to be both significantly more likely to be followed by "change talk" and significantly less likely to be followed by "sustain talk" [2]. Our finding on *Affirm* being a significant predictor of satisfaction builds upon this previous research to show that *Affirm* is also well-received by clients within an online peer-counseling platform. *Persuade (with permission)* is also found to be a significant predictor of success. While prior research on MI codes in offline mediums does not contradict this finding, this finding is also not as supported as the findings on *Reflection* and *Affirm*. Additionally,

Persuade (with permission) has a comparatively lower odds ratio of 1.019, indicating that for each added persuasive utterance, there is only 1.9% increase in the odds of the conversation being satisfactory. This finding may be especially applicable to the development of MI usage in OMHCs. In the following examples: *'I understand you must do what you think is best, but please try to eat something to stay healthy' (Listener 1)* versus *'If I can make a suggestion, you should just ignore it' (Listener 2)*, *Persuade (with permission)* clearly differs as *Listener 1* persuades the member to make a change that will directly benefit their health, whereas *Listener 2* persuades for a change that may or may not objectively benefit the member's health. These examples, combine with the statistically significant odds ratio, illustrate that *Persuade (with permission)* may be an important component of satisfying conversations on OMHCs, if listeners are trained to use this MI code effectively.

In contrast to the MI-consistent positive predictors for client satisfaction, the MI-inconsistent code *Inappropriate* has a significant negative correlation to satisfactory conversations. The significance of *Inappropriate* as a negative predictor for client satisfaction highlights the need for listener training on OMHCs. Additionally, this finding aligns with previous research that MI-consistent techniques are more successful in promoting client language supporting behavioral change during counseling sessions than MI-inconsistent techniques [2, 30]. Of further note is that the MI-consistent codes *Emphasizing Autonomy* and *Closed Question*, while not statistically significant, have low p values that are less than 0.1. These findings offer additional evidence that the success of MI-consistent technique usage translates to client satisfaction within an online medium.

6.2 RQ2: How Does Listeners' Use of Motivational Interviewing Techniques Change with Their Experience?

Now we are interested in changes in listener usage of MI strategies over time. As such, we limit the analysis to those who have been active on the platform for at least a year. This gives us the ability to look at the changes in their use of MI techniques over time, unconfounded with differences between listeners who drop out quickly from those who persist on the site. We use the same analysis technique as Section 6.1 to discover MI codes that show a change in usage as listeners gain more experience. We also examine the intersection of these identified codes with the codes from Section 6.1 that were significantly predictive of client satisfaction.

First, we query listeners who have been active on the platform. We define active listeners in two ways: (1) the time between their last and first utterances should be at least one year and (2) they have participated in at least 500 sessions, because the average number of sessions per user satisfying the first criteria is 493 (which we round to the nearest hundred). The first method gives us a temporal filter while the second method allows us to discard those listeners who are inactive and only log onto the platform after a long time. After identifying active listeners, we limit the analysis to conversations that contain at least 50 utterances to ensure that the discussion between the member and the listener moves beyond the introduction phase. After filtering the data, the average length of conversations is 713.6 utterances, with the largest containing 72,774 utterances. The relatively high average number of utterances per conversation when compared to our filter of 50 utterances shows that most conversations that exceed the threshold go on for many utterances. We use our automated labeling classifiers from Section 5 to find the MI techniques for the obtained filtered utterances. Correlational studies analysing the MI-Consistent behavior of active users is given in Appendix Subsection 8.1. The studies describe the conversational level techniques used by listeners and generalization of behavior at the individualistic listener level.

Figure 1 presents the trends in usage of motivational interviewing codes as a listener's experience increases over time. We split the behavior of the listeners into four buckets: 0 to 1 months from joining, 1 to 6 months after joining, 6 to 12 months after joining, and more than 1 year after joining. This allows us to view trends in the platform as a user gains more experience. The y axis of Figure

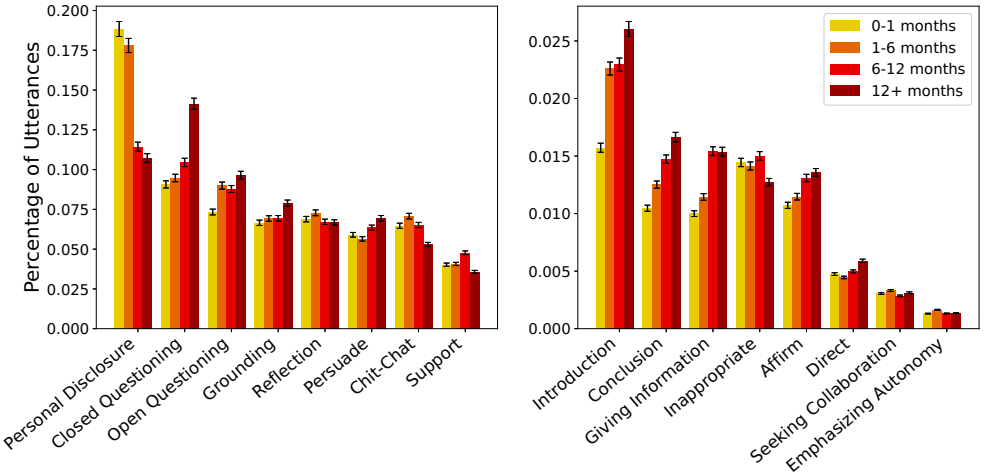

Fig. 1. Trends in Listener MI Codes over time.

1 shows the fraction of utterances for each bucket consisting of a MI technique. The left-hand graph in Figure 1 shows changes in the relatively frequent MI codes (e.g., (*Closed Question*, *Open Questions*), *Personal Disclosure*, *Grounding*, *Persuade*, *Support*, *Chit-Chat* and *Reflection*), while the right-hand graph shows changes in the less frequent MI codes (e.g., *Affirm*, *Emphasizing Autonomy* and *Seeking Collaboration*). The overall use of *Affirm* is fairly low. Since Section 6.1 showed that *Reflection* and *Affirmations* are the most useful MI technique for predicting conversation success based on Odds Ratio, these results suggest that training programs to encourage listeners to give more *Reflections* could be helpful. We see increasing trends in *Open Questions*, *Closed Questions*, *Introductions*, *Conclusions*, *Grounding*, *Affirmations*, *Persuade*, *Giving Information* and *Direct*. We also see decreasing trends in *Personal Disclosure*, *Reflection* and *Inappropriate*.

Increasing trends in *Introductions*, *Conclusions*, *Closed Questions*, and *Open Questions* show that listeners learn and refine a set method of conversing when they meet new support seekers. We also see more than 2.5% increase (given by the confidence intervals) in *Direct* and *Affirm*. This confirms that active and experienced users seem to learn to use more affirmations as they spend more time on peer-to-peer platforms. Decreasing trends are observed for *Personal Disclosure* and *Reflection*. This indicates that as listeners learn and understand the best ways to talk to support seekers, they rely less on sharing their own experiences (*Personal Disclosure*) or summarizing the issues of the support seeker (*Reflection*). The statistical significance of *Reflection* for predicting client satisfaction seen in 6.1 can be used to train listeners to learn this motivational interviewing technique.

Change in listener MI usage can be attributed to the listeners refining the way they communicate with members over time as they gain experiences. We observe this via a qualitative analysis on *Reflection*, performed on randomly sampled active listeners by comparing reflective utterances within the first month of joining and after one year of experience as a listener. Within their first month of joining as a listener, we observe utterances like: *'Haha doesn't everyone have that!' (Listener 1)*, *'That isn't very helpful.' (Listener 2)*, and *'You went to the adult side and it is a tough transition.' (Listener 3)*. These reflections are more repetitive of the member's utterance and add limited value to what members have said. On the other hand, after more than one year of experience as a listener, we see utterances which contain elements beyond paying attention; utterances which have exaggerated or amplified context, reflection of emotion/feeling that was not previously stated by member, and the use of analogies and comparatives. For example: *'I can see that he has a lot of narcissistic traits.' (Listener 1) [Amplified Context]*, *'Well, we all need a safe place and that's it. It is okay to be stressed*

about love sometimes and feel like caged birds.' (Listener 2) [Analogy - Simile], and *'I think it's easy to feel like no one really cares.' (Listener 3) [Emotion]*.

A qualitative examination of *Personal Disclosure* shows that listeners learn to be more member-centric with experience. In the beginning months of joining, listeners disclosed their own emotions, intimate experiences and circumstances, to make support seekers more comfortable: *'My parents are not supportive of college decisions.' (Listener 1)* and *'Relationships are hard for me, my last partner did not wish to spend time with me.' (Listener 2).* Alternatively, after spending more time on the platform, the listeners shared their mood, likes/ dislikes to make the member more comfortable: *'I am feeling great today!' (Listener 1)* and *'Fun fact, I am actually drinking tea right now. 10/10 I would suggest earl grey.' (Listener 2).* Similarly, examination of *Affirm* shows that listeners become more encouraging and positive as they gain more experience. For example, in the beginning months of joining, listeners only appreciate or acknowledge members' comments: *'Thank you for sharing with me' (Listener 1), 'That is an interesting perspective' (Listener 2).* After gaining experience: *'Those are amazing ideas, great work brainstorming designs' (Listener 1), 'I know it is difficult to quit smoking, I am proud of you!' (Listener 2)*; we notice distinct usage of reinforcing and complementary words. The qualitative analysis of *Affirm*, *Reflections* and *Personal Disclosures* shows that listeners adjust their interaction styles as they spend more time as a volunteer counselor.

7 CONCLUSION

This work provides a first-of-its-kind comprehensive data-set consisting of 14,797 chat messages annotated with motivational interviewing techniques on a peer-to-peer text-based volunteer counseling platform: 7Cups. Furthermore, this work translates the data into a supervised classification problem to automate the process of labeling, which helps us perform a large scale analysis of the data. Thereafter, the work focuses on a quantitative analysis which yields a few key takeaways:

- *Motivational interviewing techniques that lead to satisfactory conversations*: Our analysis shows that MI-consistent techniques are more successful in predicting client satisfaction. The techniques *Reflection*, *Affirm*, and *Persuade* have the highest statistically significant positive correlations and odds ratios to satisfactory conversations. Additionally, we find that the MI-inconsistent technique *Inappropriate* is a significant negative predictor for client satisfaction, pointing to the need for OMHCs to address inappropriate conversations in order to create better environments for both listeners and support seekers.
- *Change in behavior of listeners as they gain experience*: Active listeners learn to use the technique *Affirm* more as they gain experience. Increasing trends in *Introduction*, *Conclusion*, *Closed Question*, and *Open Question* show that listeners learn and refine a set method of conversing when they meet new members. The study also shows that listeners talk less about themselves (*Chit-Chat* and *Personal Disclosure*), which reinforces previous findings that members gain more from conversations which are member-centric [49]. We note the decrease in the usage of *Reflection* by listeners as they gain experience contrasts with its positive predictive ability of client satisfaction from Section 6.1. These findings on *Reflection* can be used to build better training programs for listeners.

7.1 Implications and Design Recommendation

Our findings build on prior MI studies by exploring MI at scale within a unique peer-counseling online medium. Similar to previous works on MI and counseling success in offline settings [2, 29, 30], we also found that MI-consistent techniques were predictive of better outcomes (defined here as client satisfaction). Our work contributes back to these prior studies by showing that MI-consistent techniques translate well to online mediums with non-professional counselors. Due to the unique

nature of the peer-counseling online medium of our study, our findings have several potential implications for practical applications within online platforms.

Previous studies have shown that listeners on platforms like 7Cups feel that while the training program prepares them to help clients, it is not sufficient in training listeners to deliver high quality counseling sessions [49]. Our work bridges this gap by providing a high-quality motivational interviewing dataset and fine-tuned language models which may be utilized for multiple cases.

We recommend the use of the dataset for building suggestive models to empower the process of support exchange in online peer counseling platforms. Given a series of chats between the member and listener, artificial intelligence models could be made to do two things. First, the models can suggest to the listener a set of MI techniques to use. Second, chatbots can suggest utterances related to the suggested MI technique to listeners. This would help new listeners, who face the issue of inadequate training [60] by guiding them to possible responses in various situations where member messages are extreme, irrational, and unreasonable. Listeners can be introduced to chatbots trained on member utterances to simulate real-life messaging with the diversity seen on peer-to-peer text-based counseling platforms. The released data can also be used for other annotations such as suicidal ideation detection. Other applications include detection of listeners who ask members for personal contact information, such listeners can be identified and flagged.

The models available can be used for real-time categorization of messages. Platforms could be integrated with inappropriate message classifiers to identify users sending obscene messages and hate speech. These users could be flagged for platform owners to review and potentially remove from the website. The classifier for *Personal Disclosure* can be used to determine the first instance when the member feels comfortable enough to disclose their simple personal information about likes/dislikes, mood, and also for more intimate information about their issues.

Lastly, the findings from our research questions give important insights into what MI techniques lead to satisfactory conversations within an online, text-based medium. Previous reports have indicated that current training programs are insufficient and leave counselors to develop their own strategies [60]. As such, this novel research can be incorporated into OMHC training programs to improve counselor confidence and success.

7.2 Limitations

This work has multiple limitations. The first limitation is that the work only covers one-to-one support seeker and peer counselor text-based platforms. It does not account for effects of group chat boards, or other Reddit-like online platforms. Second, this work does not capture the effects of offline activities, of the members and the listeners both. Many listeners connect with members on social media platforms despite being trained not to do so; this work is unable to capture interactions of users on alternate platforms. Platforms also allow volunteer counselors to seek support from other volunteer counselors, this work does not capture the impact of such interactions. It is possible that changes in MI techniques used by listeners is influenced due to the advice and guidance of other listeners. Inter-listener interactions and word-of-mouth advice may lead to drastic changes in listener behavior which are excluded from this study.

Additionally, the majority of this study is quantitative in nature and needs to be augmented with qualitative interviews from users of such platforms. Furthermore, the F1 scores for the classifier models for *Direct* and *Seeking Permission* are around 0.5; these models may contain large amounts of false negatives due to their precision heavy nature. Another limitation is while the dataset provided may help platforms to build better training programs, it does not account for extreme situations of support seeker distress/behavior. Platforms like 7Cups often have support seekers who are unresponsive to suggestions and are unwilling to communicate in an appropriate manner. Well meaning peer counselors face enormous stress and anxiety while engaging such support seekers.

This work excludes such chats from our analysis and released dataset to ensure that such support seekers do not deteriorate or relapse due to slight inaccuracy by our large language models. In such cases, peer counselors should redirect the support seekers to highly trained professional therapists and authority helpline numbers for further help.

7.3 Generalization

Our research examines motivational interviewing and client satisfaction on an OMHC. While our work is based in the specific domain of 7Cups, findings on the correlations between MI technique usage and client satisfaction may be more broadly applied to other online mental health groups.

7.4 Ethical Considerations

This research has been approved the Institutional Review Board (IRB) at our institution. All researchers involved in this study have signed strict confidentiality agreements with the 7Cups platform. As stated above, we received access to counseling session data from the platform 7Cups of Tea [40] and worked with the 7Cups' Research Advisory Board. 7Cups follows HIPAA and confidentiality agreements when collecting user data. Additionally, due to the sensitive nature of this data, we took several precautions during annotation. Annotation was not outsourced, but rather was conducted solely by the researchers working on this project, each of whom conducted the Collaborative Institutional Training Initiative (CITI Program) training prior to annotation work. Messages were also anonymized by not including the usernames of members or listeners in sessions. All examples given in the paper are paraphrased from 7Cups using round-trip translation, instead of directly pulled, to further protect user confidentiality. Additionally, to protect the anonymity of user information within our dataset, we will release our dataset and language models (Mental Health Bert) subject to appropriate privacy and ethical considerations.

ACKNOWLEDGEMENT

We would like to thank the anonymous reviewers for their helpful comments, and the team from 7 Cups for their feedback. This work is funded in part by NSF AI Institute AI-CARING (IIS-2112633), and a NSF grant IIS-2144562. DY is supported by the Microsoft Faculty Fellowship.

REFERENCES

[1] 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* 4 (2016), 463–476. https://doi.org/10.1162/tacl_a_00111

[2] Timothy R. Apodaca, Kristina M. Jackson, Brian Borsari, Molly Magill, Richard Longabaugh, Nadine R. Mastroleo, and Nancy P. Barnett. 2016. Which Individual Therapist Behaviors Elicit Client Change Talk and Sustain Talk in Motivational Interviewing? *Journal of Substance Abuse Treatment* 61 (2016), 60–65. https://doi.org/10.1016/j.jsat.2015.09.001

[3] Rita Ardito and Daniela Rabellino. 2011. Therapeutic Alliance and Outcome of Psychotherapy: Historical Excursus, Measurements, and Prospects for Research. *Frontiers in Psychology* 2 (2011). https://doi.org/10.3389/fpsyg.2011.00270

[4] D.C. Atkins, M. Steyvers, Z.E. Imel, and et al. 2014. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Sci 9* (2014). https://doi.org/10.1186/1748-5908-9-49

[5] Amit Baumel. 2015. Online emotional support delivered by trained volunteers: users' satisfaction and their perception of the service compared to psychotherapy. *Journal of Mental Health* 24, 5 (2015), 313–320. https://doi.org/10.3109/09638237.2015.1079308 arXiv:https://doi.org/10.3109/09638237.2015.1079308 PMID: 26485198.

[6] Katharina Boldt, Michaela Coenen, Ani Movsisyan, Stephan Voss, Eva Rehfuess, Angela M. Kunzler, Klaus Lieb, and Caroline Jung-Sievers. 2021. Interventions to Ameliorate the Psychosocial Effects of the COVID-19 Pandemic on Children—A Systematic Review. *International Journal of Environmental Research and Public Health* 18, 5 (2021). https://doi.org/10.3390/ijerph18052361

[7] J. K. Wood C. R. Rogers. [n. d.]. *Client-centered theory: Carl R. Rogers.* In A. Burton (Ed.), Operational theories of personality. Brunner/Mazely.

[8] Rogie Royce Carandang, Akira Shibanuma, Junko Kiriya, Karen Rose Vardeleon, Maria Aileen Marges, Edward Asis, Hiroshi Murayama, and Masamine Jimba. 2019. Leadership and Peer Counseling Program: Evaluation of Training and Its Impact on Filipino Senior Peer Counselors. *International Journal of Environmental Research and Public Health* 16, 21 (2019). https://doi.org/10.3390/ijerph16214108

[9] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for Abusive Language Detection in English. arXiv:2010.12472 [cs.CL]

[10] Prerna Chikersal, Danielle Belgrave, Gavin Doherty, Angel Enrique, Jorge E. Palacios, Derek Richards, and Anja Thieme. 2020. Understanding Client Support Strategies to Improve Clinical Outcomes in an Online Mental Health Intervention. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376341

[11] Joe Curran, Glenys Parry, Gillian Hardy, Jennifer Darling, Ann-Marie Mason, and Eleni Chambers. 2019. How does therapy harm? A model of adverse process using task analysis in the synthesis of service users' experience. *Frontiers in Psychology* (03 2019). https://doi.org/10.3389/fpsyg.2019.00347

[12] Simon D'Alfonso, Olga Santesteban-Echarri, Simon Rice, Greg Wadley, Reeva Lederman, Christopher Miles, John Gleeson, and Mario Alvarez-Jimenez. 2017. Artificial Intelligence-Assisted Online Social Therapy for Youth Mental Health. *Frontiers in Psychology* 8 (2017), 796. https://doi.org/10.3389/fpsyg.2017.00796

[13] Larry Davidson. 2013. Peer Support: Coming of Age of and/or Miles to Go Before We Sleep? An Introduction. *The journal of behavioral health services and research* 42 (11 2013). https://doi.org/10.1007/s11414-013-9379-2

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[15] Nikolaos Flemotomos, Victor R. Martinez, Zhuohao Chen, Karan Singla, Victor Ardulov, Raghuveer Peri, Derek D. Caperton, James Gibson, Michael J Tanana, Panayiotis G. Georgiou, Jake Van Epps, Sarah Peregrine Lord, Tad Hirsch, Zac E. Imel, David C. Atkins, and Shrikanth S. Narayanan. 2021. "Am I A Good Therapist?" Automated Evaluation Of Psychotherapy Skills Using Speech And Language Technologies. *Behavior research methods* (2021).

[16] Ruben Fukkink. 2011. Peer Counseling in an Online Chat Service: A Content Analysis of Social Support. *Cyberpsychology, behavior and social networking* 14 (04 2011), 247–51. https://doi.org/10.1089/cyber.2010.0163

[17] Katy Ilonka Gero and Lydia B. Chilton. 2019. *Metaphoria: An Algorithmic Companion for Metaphor Creation.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300526

[18] Isabel Gibbard and Terry Hanley. 2008. A five-year evaluation of the effectiveness of person-centred counselling in routine clinical practice in primary care. *Counselling and Psychotherapy Research* 8, 4 (2008), 215–222. https://doi.org/10.1080/14733140802305440 arXiv:https://doi.org/10.1080/14733140802305440

[19] Louis F Graham, Halley P Crissman, Jack Tocco, Laura A Hughes, Rachel C Snow, and Mark B Padilla. 2014. Interpersonal Relationships and Social Support in Transitioning Narratives of Black Transgender Women in Detroit. *International Journal of Transgenderism* 15 (2014), 100–113.

[20] Z. E. Imel, B. T. Pace, C. S. Soma, M. Tanana, T. Hirsch, J. Gibson, P. Georgiou, S. Narayanan, and D. C. Atkins. 2019. Design feasibility of an automated, machine-learning based feedback system for motivational interviewing. *Psychotherapy* (2019), 318–328. https://doi.org/10.1037/pst0000221

[21] J. Kaiser, F. Hanschmidt, and A Kersting. 2021. The association between therapeutic alliance and outcome in internet-based psychological interventions: A meta-analysis. *Computers in Human Behavior* (2021). https://doi.org/10.1016/j.chb.2020.106512

[22] Neel Kant, Raul Puri, Nikolai Yakovenko, and Bryan Catanzaro. 2018. Practical Text Classification With Large Pre-Trained Language Models. *CoRR* abs/1812.01207 (2018). arXiv:1812.01207 http://arxiv.org/abs/1812.01207

[23] Wang PS Kessler RC. 2008. The descriptive epidemiology of commonly occurring mental disorders in the United States. *Annu Rev Public Health* (2008). https://doi.org/10.1146/annurev.publhealth.29.020907.090847

[24] Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. (2011).

[25] Taisa Kushner and Amit Sharma. 2020. *Bursts of Activity: Temporal Patterns of Help-Seeking and Support in Online Mental Health Forums.* Association for Computing Machinery, New York, NY, USA, 2906–2912. https://doi.org/10.1145/3366423.3380056

[26] Fei-Tzin Lee, Derrick Hull, Jacob Levine, Bonnie Ray, and Kathy McKeown. 2019. Identifying therapist conversational actions across diverse psychotherapeutic approaches. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology.* Association for Computational Linguistics, Minneapolis, Minnesota, 12–23. https://doi.org/10.18653/v1/W19-3002

[27] TalkLife Limited. 2021. *TalkLife.* https://www.talklife.com/

[28] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[29] Brad Lundahl, Chelsea Kunz, Cynthia Brownell, Derrik Tollefson, and Brian Burke. 2010. A Meta-Analysis of Motivational Interviewing: Twenty-Five Years of Empirical Studies. *Research on Social Work Practice - RES SOCIAL WORK PRAC* 20 (03 2010), 137–160. https://doi.org/10.1177/1049731509347850

[30] Magill M, Gaume J, Apodaca TR, Walthers J, Mastroleo NR, Borsari B, and Longabaugh R. 2014. The technical hypothesis of motivational interviewing: a meta-analysis of MI's key causal model. *J Consult Clin Psychol* (12 2014), 973–83. https://doi.org/10.1037/a0036833

[31] Davis MK Martin DJ, Garske JP. 2000. Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review. *J Consult Clin Psychol* (2000), 438–450. https://doi.org/10.1037/0022-006X.68.3.438

[32] Petra S Meier, Christine Barrowclough, and Michael C Donmall. 2005. The role of the therapeutic alliance in the treatment of substance misuse: a critical review of the literature. *Addiction* 100, 3 (2005), 304–316.

[33] MellowTalk. 2021. *MellowTalk.* https://mellowtalk.com/

[34] William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2003. *Manual for the motivational interviewing skill code (misc). Unpublished manuscript.* Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico.

[35] Rollnick S Miller WR. 2013. *Motivational Interviewing: Helping People Change.* Guilford Press.

[36] T.B. Moyers, J.K. Manuel, and D Ernst. 2014. Motivational Interviewing Treatment Integrity Coding Manual 4.1. (2014).

[37] T.B. Moyers, J.K. Manuel, and D. Ernst. 2014. *Motivational Interviewing Treatment Integrity Coding Manual 4.1. Unpublished manual.* University of New Mexico Center on Alcoholism, Substance Abuse, and Addictions (CASAA).

[38] J A Naslund, K A Aschbrenner, L A Marsch, and S J Bartels. 2016. The future of mental health care: peer-to-peer support and social media. (2016), 113–122. https://doi.org/10.1017/S2045796015001067

[39] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.* 9–14.

[40] 7 Cups of Tea. 2021. *7cups.* https://www.7cups.com/

[41] Zhenhui Peng, Qingyu Guo, Ka Wing Tsang, and Xiaojuan Ma. 2020. *Exploring the Effects of Technological Writing Assistance for Support Providers in Online Mental Health Community.* Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376695

[42] Verónica Pérez-Rosas, Xuetong Sun, Christy Li, Yuchen Wang, Kenneth Resnicow, and Rada Mihalcea. 2018. Analyzing the Quality of Counseling Conversations: the Tell-Tale Signs of High-quality Counseling. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1591

[43] Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What Makes a Good Counselor? Learning to Distinguish between High-quality and Low-quality Counseling Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, Florence, Italy, 926–935. https://doi.org/10.18653/v1/P19-1088

[44] Yada Pruksachatkun, Sachin R. Pendse, and Amit Sharma. 2019. Moments of Change: Analyzing Peer-Based Cognitive Support in Online Mental Health Forums. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019).

[45] Brian Rodgers, Keith Tudor, and Anton Ashcroft. 2021. Online video conferencing therapy and the person-centered approach in the context of a global pandemic. *Person-Centered & Experiential Psychotherapies* 20, 4 (2021), 286–302. https://doi.org/10.1080/14779757.2021.1898455 arXiv:https://doi.org/10.1080/14779757.2021.1898455

[46] Koustuv Saha, Sindhu Kiranmai Ernala, Sarmistha Dutta, Eva Sharma, and Munmun De Choudhury. 2020. Understanding Moderation in Online Mental Health Communities. In *Social Computing and Social Media. Participation, User Experience, Consumer Experience, and Applications of Social Computing*, Gabriele Meiselwitz (Ed.). Springer International Publishing, Cham, 87–107.

[47] Koustuv Saha and Amit Sharma. 2020. Causal Factors of Effective Psychosocial Outcomes in Online Mental Health Communities. *Proceedings of the International AAAI Conference on Web and Social Media* 14, 1 (May 2020), 590–601. https://ojs.aaai.org/index.php/ICWSM/article/view/7326

[48] SAMHSA - Substance Abuse and Mental Health Services Administration. 2013. TIP 42: Substance Abuse Treatment For Persons With Co-Occurring Disorders : A Treatment Improvement Protocol. https://store.samhsa.gov/system/files/sma13-3992.pdf

[49] Haard Shah, Robert Kraut, Yi-Chia Wang, and Diyi Yang. 2022. Understanding the Effectiveness of Peer2Peer Counseling with Multiple Feedback Channels. *Unpublished Manuscript* (2022). https://doi.org/10.1145/1122445.1122456

[50] Diener M. J Sharf J., Primavera L. H. 2010. Dropout and therapeutic alliance: A meta-analysis of adult individual psychotherapy. *Psychotherapy: Theory, Research, Practice, Training* (2010), 637–645. https://doi.org/10.1037/a0021175

[51] Eva Sharma and Munmun De Choudhury. 2018. *Mental Health Support and Its Relationship to Linguistic Accommodation in Online Communities.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/

3173574.3174215
[52] P. Solomon. 2004. *Peer Support/Peer Provided Services Underlying Processes, Benefits, and Critical Ingredients*. Vol. 27(4). 392–401. https://doi.org/10.2975/27.2004.392.401
[53] Keith Tudor and David Murphy. 2021. Online therapies and the person-centered approach. *Person-Centered & Experiential Psychotherapies* 20, 4 (2021), 283–285. https://doi.org/10.1080/14779757.2021.2000139 arXiv:https://doi.org/10.1080/14779757.2021.2000139
[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR* abs/1706.03762 (2017). arXiv:1706.03762 http://arxiv.org/abs/1706.03762
[55] Liuping Wang, Xiangmin Fan, Feng Tian, Lingjia Deng, Shuai Ma, Jin Huang, and Hongan Wang. 2018. MirrorU: Scaffolding Emotional Reflection via In-Situ Assessment and Interactive Feedback *(CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3170427.3188517
[56] Street N Wang X, Zhao K. 2017. Analyzing and Predicting User Participations in Online Health Communities: A Social Support Perspective. *J Med Internet Res 2017;19(4):e130* (2017). https://doi.org/10.2196/jmir.6834
[57] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *CoRR* abs/1910.03771 (2019). arXiv:1910.03771 http://arxiv.org/abs/1910.03771
[58] Shaomei Wu, Lindsay Reynolds, Xian Li, and Francisco Guzmán. 2019. *Design and Evaluation of a Social Media Writing Support Tool for People with Dyslexia.* Association for Computing Machinery, New York, NY, USA, 1–14.
[59] Diyi Yang, Robert Kraut, and John M. Levine. 2017. Commitment of Newcomers and Old-Timers to Online Health Support Communities. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6363–6375. https://doi.org/10.1145/3025453.3026008
[60] Zheng Yao, Robert Kraut, and Haiyi Zhu. 2018. Learning to Become a Volunteer Counselor: Lessons from a Peer-to-Peer Mental Health Community. *Under Review* (2018). https://doi.org/10.1080/14733140802305440

8 APPENDIX

8.1 Correlational analysis on the behaviors of active users

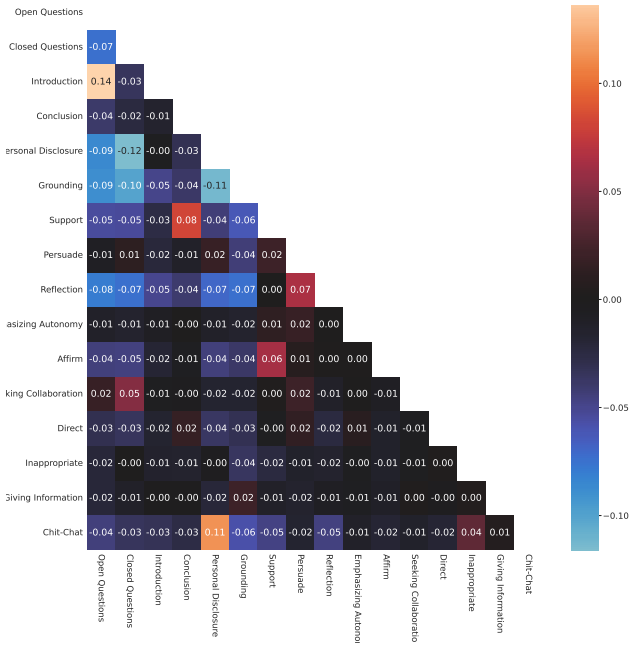

Fig. 2. Correlation matrix for co-occurence of MI Codes

We begin our correlational analysis on active users by finding which MI codes co-occur at an utterance level to understand listeners who are invested into online counseling platforms. The correlation matrix in figure 2 is found on the multi hot encoding of the predicted codes. The overall correlation values in the figure are very low, although we observe noticeable positive correlations between *Introduction* and *Open Questions*, which is expected as discussed in Section 1. We do not pursue this research direction to find the efficacy of multiple MI codes in the same utterance because users can send multiple messages in the same turn. Users treat an utterance as a unit of text which generally conveys one idea. The other reason lies in the fact that some MI codes are used much more than others, which pushes the correlation between codes towards negative values due to the multi hot encoding. This approach focuses on only the co-occurrence of MI codes, for our next analysis, we focus on the conversation level techniques used by our listeners. At a conversational level

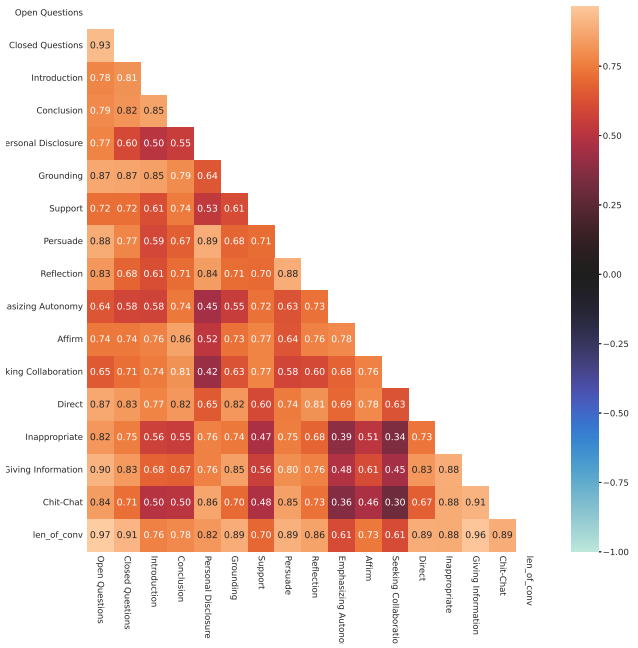

Fig. 3. Co-relational Analysis at the Conversation Level

(figure 3), we find that as the length of the conversation increases, the proportion of MI-consistent codes like *Affirm*, *Emphasizing Autonomy*, and *Seeking Collaboration* decreases (i.e. as the length of the conversation increases, the counts of the MI-consistent codes does not increase with the same rate). The other finding is that *Inappropriate* has some of the lowest correlation values in the entire matrix with all the other MI codes except for *Giving Information*. This implies that active listeners are on the platform to spread positive messages. The conversation level correlation analysis seems to show that *Personal Disclosure* and *Inappropriate* have lower co-relation with the MI-consistent codes like *Affirm*, *Seeking Collaboration*, *Emphasizing Autonomy etc.* This indicates that listeners should not focus the conversations significantly on themselves.

The third type of correlational analysis focuses on the generalization of behaviors at the individualistic listener level. Co-relation between MI codes at the listener level would help us find possible relations such as a listener that sends more inappropriate messages may also send more messages to persuade the members. We find the co-relational matrix for the listener level in figure

4. We observe a high correlation between reflection and persuasion. This implies that listeners often repeat and reflect on what the member has uttered, and they also persuade the users to take up action items. Positive correlations between (*Seeking Collaboration* and *Conclusion*), and (*Affirm* and *Conclusion*) show that experienced listeners tend to take their conversations to its conclusion before signing off. Other positive and negative listener level co-relation analysis show that in most cases experienced listeners follow MI-consistent behavior.

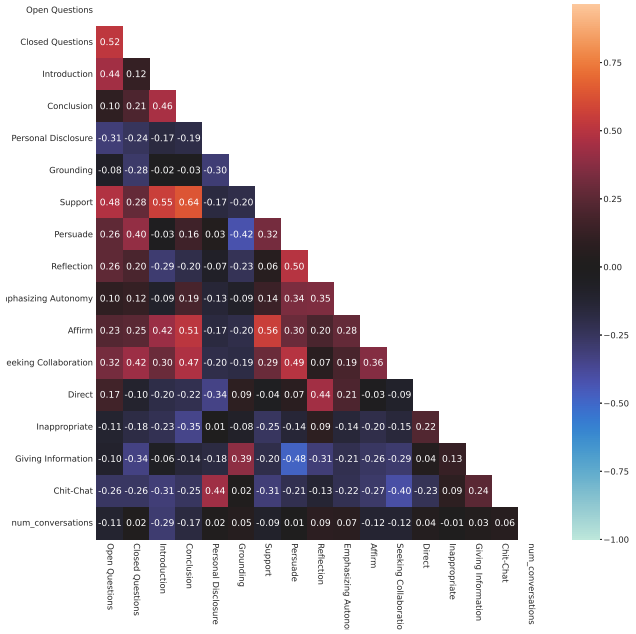

Fig. 4. Co-relational Analysis at the Listener Level

Received January 2022; revised April 2022; accepted August 2022

Architectures:	BERT-Base-Uncased (F1 Scores)			BERT-large (F1 Scores)		
MI Codes \ Context:	None	1 previous utterance	5 previous utterances	None	1 previous utterance	5 previous utterances
Giving Information	0.356	0.329	0.177	0.422	0.371	0.318
Reflection	0.567	0.543	0.608	0.574	0.607	0.658
Support	0.630	0.633	0.671	0.608	0.630	0.711
Affirm	0.485	0.425	0.333	0.462	0.397	0.526
Closed Question	0.894	0.889	0.858	0.890	0.877	0.884
Persuade	0.595	0.568	0.649	0.602	0.740	0.669
Open Questions	0.899	0.904	0.889	0.904	0.916	0.904
Seeking Collaboration	**0.353**	0.368	NA	0.406	0.388	NA
Inappropriate*	0.342	0.374	0.261	0.517	0.454	0.440
Direct	0.394	0.311	0.087	0.410	0.419	0.182
Emphasizing Autonomy	0.629	0.621	NA	0.694	0.600	NA
Grounding	0.724	0.743	0.772	0.749	0.814	0.803
Personal Disclosure	0.731	0.719	0.790	0.771	0.745	0.804
Introduction	0.854	0.897	0.730	0.871	0.894	0.800
Conclusion	0.716	0.729	0.413	0.701	0.732	0.465
Chit-Chat	0.178	0.619	0.193	0.293	0.660	0.240
Other	0.345	0.539	0.667	0.412	0.622	0.681
Avg F1 Scores (Across MI Codes)	0.570	0.600	0.539	0.605	0.639	0.608

Table 7. F1 scores of the Positive Classes for MI Codes using 0, 1, and 5 utterances as previous context when using BERT-Base-Uncased and BERT-Large architectures without domain specific pre-training.

Architectures:	BERTweet (F1 Scores)			Roberta-Base (F1 Scores)		
MI Codes \ Context:	None	1 previous utterance	5 previous utterances	None	1 previous utterance	5 previous utterances
Giving Information	0.422	0.517	0.233	0.385	0.410	0.293
Reflection	0.585	0.705	0.727	0.580	0.650	0.649
Support	0.692	0.690	0.775	0.656	0.651	0.677
Affirm	0.672	0.542	0.327	0.621	0.512	0.387
Closed Question	0.901	0.901	0.862	0.896	0.896	0.894
Persuade	0.605	0.747	0.816	0.584	0.616	0.765
Open Questions	0.911	0.930	0.921	0.905	0.915	0.914
Seeking Collaboration	0.512	0.465	NA	0.439	0.359	NA
Inappropriate*	0.441	0.595	0.612	0.412	0.459	0.462
Direct	0.492	0.500	0.276	0.415	0.400	0.182
Emphasizing Autonomy	0.720	0.667	NA	0.769	0.533	NA
Grounding	0.776	0.822	0.828	0.738	0.767	0.810
Personal Disclosure	0.791	0.755	0.827	0.776	0.741	0.800
Introduction	0.895	0.908	0.841	0.855	0.893	0.828
Conclusion	0.798	0.748	0.521	0.739	0.710	0.489
Chit-Chat	0.269	0.680	0.397	0.253	0.604	0.214
Other	0.446	0.623	0.769	0.310	0.567	0.755
Avg F1 Scores (Across MI Codes)	0.642	0.693	0.649	0.608	0.628	0.606

Table 8. F1 scores of the Positive Classes for MI Codes using 0, 1, and 5 utterances as previous context when using vinai/BERTweet and Roberta-Base architectures without domain specific pre-training.

Class Name (Abbreviation)	Inter Annotator Agreement	Agreement Between the model and the Annotators
Giving Information (GI)	1	0.663
Reflection (RF)	0.587	0.552
Support (SUP)	0.710	0.594
Affirm (AFFIRM)	1	0.795
Closed Question (QUC)	0.826	0.826
Open Question (QUO)	0.808	0.729
Persuade (PR)	0.581	0.47
Direct (DI)	1	0.663
Emphasizing Autonomy (AUTO)	1	1
Grounding (GR)	0.906	0.796
Personal Disclosure (PD)	0.594	0.499
Introduction/Greeting (INT)	1	0.795
Outro/Conclusion (OUT)	0.795	0.795
Seeking Collaboration/Permission (SEEK)	1	0.663
Inappropriate (IP)	NAN	NAN
Chit-Chat (CC)	0.71	0.426
Cumulative Agreement	0.734	0.614

Table 9. Validating the Binary classifiers by calculating Krippendorff Alphas for between the Annotators and the MH-BERT Model.