Detecting Change Talk in Motivational Interviewing using Verbal and Facial Information

Yukiko I Nakano, Jean-Claude Martin, Tatsuya Sakato, Shogo Okada, Jean-Claude Martin

Detecting Change Talk in Motivational Interviewing using Verbal and Facial Information

Yukiko I. Nakano
Seikei University
Musashino-shi, Tokyo, Japan
y.nakano@st.seikei.ac.jp

Eri Hirose
Seikei University
Musashino-shi, Tokyo, Japan
dm216216@cc.seikei.ac.jp

Tatsuya Sakato
Seikei University
Musashino-shi, Tokyo, Japan
sakato@st.seikei.ac.jp

Shogo Okada
Japan Advanced Institute of Science and Technology (JAIST)
Nomi, Ishikawa, Japan
okada-s@jaist.ac.jp

Jean-Claude Martin
Université Paris-Saclay, CNRS
Paris, France
jean-claude.martin@universite-paris-saclay.fr

ABSTRACT

Behavior change is one of the most important goals in psychotherapy. This study focuses on Motivational Interviewing (MI), which is collaborative communication aimed at eliciting the client's own reasons for behavior change. To investigate the effectiveness of facial information in modeling MI, we collected an MI encounter corpus with speech and video data in the nutrition and fitness domains and annotated client utterances using the Manual for the Motivational Interviewing Skill Code (MISC). By analyzing client answers to the questions after the session, we found that clients who expressed more Change Talk were more motivated to change their behavior than those who expressed less Change Talk. We then proposed RNN-based multimodal models to detect Change Talk by setting a 2-class classification task: "Change Talk" and "not Change Talk." Our experiment showed that the best performing model was a multimodal BiLSTM model that fused language and client facial information. We also found that fusing language and facial information as context achieved better performance than the unimodal and no-context models. Moreover, we discuss the label imbalance problem and conduct an additional analysis using turns as a unit of analysis. As a result, our best model reached F1-score of 0.65 for Change Talk detection.

CCS CONCEPTS

• **Computing methodologies** → *Artificial intelligence*; *Neural networks*; • **Human-centered computing** → *Empirical studies in HCI*; *Interaction techniques.*

KEYWORDS

motivational interviewing; change talk; facial information; multimodal neural networks

ICMI '22, November 7–11, 2022, Bengaluru, India

ACM ISBN 978-1-4503-9390-4/22/11…$15.00
https://doi.org/10.1145/3536221.3556607

ACM Reference Format:
Yukiko I. Nakano, Eri Hirose, Tatsuya Sakato, Shogo Okada, and Jean-Claude Martin. 2022. Detecting Change Talk in Motivational Interviewing using Verbal and Facial Information. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3536221.3556607

1 INTRODUCTION

Not having a healthy lifestyle (nutrition or physical activity) is dangerous for your health. However, it is difficult to change these behaviors. Many theories and models [16] as well as different therapeutic techniques such as Motivational Interviewing (MI) have been proposed in psychology. MI was initially proposed to help alcoholics, but since then, it has been applied and adapted to multiple domains such as nutrition and fitness [6]. Motivational Interviewing is defined as *"a collaborative, goal-oriented style of communication with particular attention to the language of change; it is designed to strengthen personal motivation for and commitment to a specific goal by eliciting and exploring the person's own reasons for change within an atmosphere of acceptance and compassion."* [18]. Thus, the primary goal of MI is to elicit the client's Change Talk (CT) by expressing willingness to change through dialogue. The professional is called the "counselor". The person who needs to change his or her behavior is called the "client".

Techniques that seek to authoritatively impose a behavior change do not work. Instead, MI relies on very subtle communication from the therapist with different components, including encouraging CT, that is, the therapist should try, through dialogue, to get the patient/client to talk more about change (and not about the current unhealthy situation). To train such MI skills, the Motivational Interviewing Skill Code (MISC) [17] was developed to evaluate the quality of MI. MISC provides categories of client and counselor utterances, and CT is one of the client utterance categories. MI trainers and trainees themselves analyze audio and video of counseling sessions using MISC. However, manual analysis of these data is time-consuming and costly. Therefore, automatic utterance classification in MI is expected to support MI skill training. For instance, it supports MI trainers to give advises to trainees in effective and efficient ways. It also helps professional counselors better understand how client behaviors change through therapy sessions.

With a similar motivation, many studies have proposed machine learning models that classify client and counselor utterances. Most of these studies used only language information [4, 5, 11, 21, 23, 26, 32] and a few studies used speech and language [1, 25]. However, little has been studied whether facial features contribute to the creation of better MI models. The literature on MI notes the importance of nonverbal cues [6]. While clients may verbally express that they are ready to make a behavior change, nonverbal cues may express consistent or ambivalent signs. For example, a client might say, "Sure, I guess I could try that." But she may also produce negative facial expressions [6]. Therefore, it is expected that integrating facial information is an effective way to analyze and interpret MI encounters.

Based on the discussion above, this study aims to detect CT using language and facial information by employing multimodal machine learning techniques. We focus on CT because this is the most important sign that a therapist wants to elicit from the client. We propose 2-class classification models that classify CT and not-CT using language and facial features. Furthermore, we conduct an experiment to examine whether facial information is useful in detecting CT. The contributions of this study are as follows.

- Collecting speech and video data for client-counselor interaction in MI sessions
- Creating an MI corpus with client utterance category annotation
- Analyzing the relationship between CT and client motivation after MI
- Creating RNN-based multimodal models that detect CT

Furthermore, the main findings of our model evaluation experiment are as follows: (1) multimodal models that fuse language and facial information along with preceding utterances as context achieved better performance compared with unimodal models and models without context. (2) we will discuss how the label imbalance problem impacts model performance by comparing it with previous studies, and (3) we propose a turn-based evaluation method based on examining whether CT is the main intention of the current client turn.

The following section describes previous studies on MI and social signal processing. Section 3 describes the MI session recording and collected data. Section 4 proposes a method for detecting CT, and Section 5 describes our experimental results and discussion. Finally, in Section 6, limitations and future work are discussed.

2 BACKGROUND

2.1 MI and multimodal interaction

The community of MI practitioners is highly structured, with international associations promoting MI and representative associations in different countries. These associations aim to train interested practitioners in this technique and keep a technological watch on techniques that may be of interest to MI. In addition, several studies in multimodal Human-Computer Interaction research have designed virtual agents that take on the role of coaches, proposing a motivational interview with the user [14]. Whether to inform the design of these virtual coaches or to better understand the MI technique to train it properly, it would be interesting to analyze videos of MI interviews automatically.

Motivational Interviewing is grounded in a respectful stance, focusing on building rapport in the initial stages of the counseling relationship. CT is defined as statements by the client revealing consideration of, motivation for, or commitment to change. Thus, in MI, the therapist seeks to guide the client to the expression of CT as the pathway to change. The different types of CT include: Desire (I want to change), Ability (I can change), Reason (It's important to change), Need (I should change), Commitment (I will make changes), Taking Steps (I am taking specific actions to change). Research indicates a correlation between client statements about change and outcomes when the client reported levels of success in changing a behavior[27].

The MI technique has evolved over the years. For example, it is now recognized that the therapist should spend more time talking about the goal in terms of behavior change (CT) than trying to have the client talk about his or her current unhealthy behaviors. Language is, therefore, one of the most important communication modalities in MI, since one of the therapists' goals is to elicit a discourse about change from the client. However, as in any dyadic interaction between a health professional and a patient, other modalities such as facial expressions and gaze are important to establish rapport with the client, to encourage him/her to speak, or to acknowledge her change talk. For example, it was observed that patients who reported being ambivalent about behavior change displayed blended expressions of positive and negative emotions [12].

2.2 Social Signal Processing in Healthcare

Many previous studies have focused on multimodal modeling for healthcare applications. In particular, multimodal behavioral sensing and multimodal machine learning techniques help develop recognition models of physiological aspects or cognitive states as social signals [2, 29] for healthcare applications. For healthcare applications, depression analysis [20], dementia detection [19], and patient pain detection [30] have been studied. According to [20], facial features, including facial expression, gaze activity, and head motion, are important descriptors for detecting a depressed state during interviewing. We hypothesized that the client's facial features are effective in capturing the client's inner state, and these features are key descriptors as a trigger for CT.

Several studies have analyzed clients' multimodal behaviors in interviewing or counseling in human-human and human-agent settings. Xiao et al. [33] presented an automatic analysis model of empathy with behavioral signal processing [33] to implement counselor agents that are used for physiological healthcare.

Several studies have focused on analyzing counselor behavior. Professional counselors are required to listen to patients' stories empathetically. Therefore, analyzing multimodal behaviors to represent empathy is an important challenge. DeVault et al. [7] presented a virtual human interviewer system designed to have face-to-face interactions in which the user feels comfortable talking and sharing information. The key technique is to adapt the agent's nonverbal behavior based on recognizing the multimodal behavior of users, including facial expressions and acoustic features [22]. In particular, the system in [7] was designed to create interactive situations favorable for the automatic assessment of distress indicators, defined as

verbal and nonverbal behaviors correlated with depression, anxiety, or posttraumatic stress disorder (PTSD). Using a corpus for user-agent interaction [7], Tavabi et al. [24] analyzed behavioral cues that indicate an opportunity to provide an empathetic response using a multimodal deep neural network. As studies focusing on MI, Xiao et al. [34] proposed a prediction model for counselor empathy measures in motivational interviewing. Wu et al. [31] proposed a turn-level detection model for client needs for empathy by leveraging pre-trained language models and empathy-related general conversation corpora.

Studies more closely related to this study classified client utterances into three classes: Change Talk (CT), in which the client is ready to change; Sustain Talk (ST), expressing resistance to change; and Follow/Neutral (FN), which is a statement unrelated to change. As studies proposed multimodal models, Aswamenakul et al. [1] proposed a logistic regression model to classify the three types of client utterances using acoustic and language features. Tavabi et al. [25] also analyzed linguistic and acoustic features and developed a neural network using pre-trained models, BERT and VGGish, to obtain language and speech embeddings. They also used the GRU to encode speech information. Other studies have proposed client utterance classification models using only language information by employing different learning algorithms, such as CRF [4], RNN [23], GRU [26, 32], and LSTM with attention mechanism [11]. Cao et al. [5] proposed a GRU-based dialogue encoder with word- and sentence-level attention. However, no study has used facial features to classify client/counselor utterances. Therefore this study investigate whether facial features contribute to detecting client CT.

It is also notable that in [11, 23, 32], only target utterances for classification were input to the network, and [5, 25, 26] used preceding utterances as the context. However, none of them exploited nonverbal information in preceding utterances. We aim to create multimodal models that effectively exploit both verbal and nonverbal context.

3 CORPUS COLLECTION

In order to collect a corpus containing visual information, we recorded motivational interview sessions.

3.1 Topics and participants

Four clinicians with professional MI skills participated in the study as counselors. They were psychotherapists and healthcare professionals. The recording sessions were not conducted in a medical setting for the purpose of our corpus collection. Each counselor participated in 12 to 13 sessions. For the clients, we recruited 52 people who wanted to improve their diet: controlling food intake (overeating); difficulty in having a well-balanced diet; controlling excessive intake of salt, fat, and sugar; solving picky eating problems, etc. The average age of the clients was 35 years old; 27 participants were male and 25 were female. We focused on diet because many people are concerned with diet in their daily lives, and we expected that they would be fully engaged in counseling encounters. All counselors and clients were native Japanese speakers, and the corpus was collected in Japanese. This study was approved by the ethical review committee, and data were collected with the consent of the participants.

3.2 Recording environment

Counseling encounters were conducted using a Zoom remote communication tool to prevent the spread of COVID-19. The counselors participated in the sessions at their individual sites. We tested whether their environment, including the Internet connection, was suitable for conducting and recording sessions. The clients were asked to come to a recording booth so that the recording environment for all clients was the same. For both sides of the participants, only their partner's upper body was shown on the display, in full-screen mode. Their own images were not shown on the display. An experimenter hosted a Zoom meeting and another staff member joined the meeting as a co-host. Therefore, remote communication during the session was stable even if the network connection at the host site was unstable.

3.3 Procedure

When both the counselor and client joined the meeting, we first confirmed that their audio and video were sufficiently clear and set Zoom to show the partners' images in full-screen mode. After confirming that the client had already thought about what they would like to consult, the experimenter asked them to start a 20 min counseling session.

3.4 Data

3.4.1 Recording. Using the recording function in Zoom, we obtained separate recordings for the counselor and client audio in addition to the mixed audio for both the counselor and client. However, Zoom does not provide individual video recordings, and we used Zoom video recording only for the counselor (the videos contained audio for both the counselor and client). Client videos were recorded using another camera placed in front of their faces. Thus, the client recorded image was almost the same as what the counselor viewed during video communication. Because we collected multiple audio and visual sources, we synchronized all sources with respect to the timestamps in the client videos. A voice activity detection (VAD) program was applied to the individual audio for counselors and clients to detect speech intervals. The sampling frequency was set to 32,000 Hz and the threshold for the amplitude level was 400. By applying this threshold, when a silent interval longer than 200 ms was detected, silence was identified as the utterance boundary. These individual audio sources were processed using Google ASR to obtain transcripts. Since we used separate audio recordings for counselor and client speech, we did not have a problem with overlapping speech. When severe speech recognition errors were found, they were corrected manually. Four sessions experienced recording problems such as missing videos, and we used 48 sessions for further analysis. From the 48 MI counseling sessions, 12,346 client and 11,297 counselor utterances were obtained. The average length of a client utterance was 2.33 s, and that of a counselor utterance was 2.75 s. The average duration of the sessions was 21 min 57 s.

3.4.2 Corpus Annotation using MISC. We annotated the transcribed utterances using the Manual for the Motivational Interviewing Skill Code (MISC) version 2.1 [17]. The MISC is a coding scheme used to analyze MI sessions worldwide. The MISC provides a coding scheme for both counselor and client utterances. For client utterance coding, utterance categories such as Reason, Taking-Steps, Commitment, and Follow/Neutral (FN) were defined as introduced in Section 2. The Reason category was then decomposed into subcategories. Except for Follow/Neutral, these categories have positive (+) and negative (-) valence. When a client utterance expressed an inclination toward change, such as Reason+ and Taking-Steps+, these utterances were identified as CT. In contrast, when a client utterance expressed moving away from change, such as Reason- and Taking-Steps-, the utterance was identified as Sustain Talk (ST). From this process, all client utterances were labeled as ST, CT, or FN. We calculated the inter-coder reliability between two annotators for the three-class categorization, and Cohen's Kappa value was 0.64 (substantial agreement).

Table 1 shows an interaction example in our corpus with its MISC code. In this example, first, the client explained his eating habit as an FN. He then discussed why he needed to stop this habit as CT. However, right after this utterance, he expressed his ambivalent feeling by declaring that this habit was hard to stop. As shown in this example, we frequently observed a sequence of short utterances by the same speaker. This is because we used VAD, which was automatically identified as a unit of analysis and was called it an utterance.

Table 2 shows the distribution of the client utterances across the three categories. 76 % of the client utterances were FN, and only 15 % were CT. Thus, the distribution of categories was imbalanced. This is the first encounter between the client and the counselor, and the early part of the conversation (the first 30 utterances) was spent on greetings, asking each other's names, and the client's brief explanation of her/his diet problem. As these utterances were not included in the body of the MI, we excluded this part from further analysis[1].

3.4.3 Post session questionnaire. After the session, the clients were requested to answer a questionnaire that included a question about the subjective evaluation of their current motivation level. The question was, "When will you change your behavior? The choices were "today," "tomorrow," "this week," "this month," "not sure," and "cannot change."The distribution of the answers from the 48 clients is shown in Table 3. Some clients answered that they were thinking about changing their behavior soon, and some were unmotivated.

We examined whether clients' behavior during MI affected their feelings after the session. We divided the clients into two groups. The 25 clients who answered "today" or "tomorrow" were assigned to a high motivation(HM) group; the 13 clients who answered "this month," "not sure," or " cannot change" were assigned to a low motivation (LM) group.

As shown in Figure 1, the average number of CT was higher in the HM group than that in the LM group. The difference was statistically significant in t-test ($t(36) = 2.028$, $p = 0.05$, Cohen's $d = 0.69$). Cohen's d value was 0.69; thus, the effect size was medium. This result suggests that clients were more motivated when they uttered more CT. This also supports our claim that CT detection is useful for estimating the client's motivation level (at least) right after the MI encounter.

[1]Table 2 shows the numbers before excluding the first 30 utterances

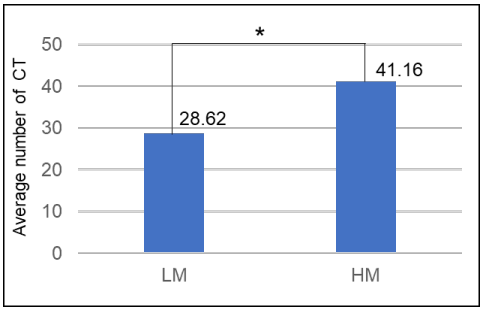

Figure 1: Average number of CT in LM and HM groups. * indicates the difference is statistically significant at the 0.05 level (p<=0.05) in t-test.

4 DETECTING CHANGE TALK

In this section, we propose a method for detecting CT in MI sessions. Previous studies used only language information [5, 26] or fused language and speech information to create client utterance classification models and reported that speech information did not contribute to improving model performance [25]. In this study, we introduced the facial features of the participants in detecting CT and investigated whether model performance improves by combining language and visual information.

First, we describe feature extraction and then propose neural network models that use these features as inputs.

4.1 Feature extraction

4.1.1 Language features. We used RoBERTa [15], which improves the end-task performance by modifying the BERT pre-training procedure. The difference from BERT is that RoBERTa trains the model longer, with larger batches over more data, removing the next sentence prediction objective, training on longer sequences, and dynamically changing the masking pattern applied to the training data. A previous study reported that, in client utterance classification in MI, RoBERTa outperformed BERT[8]. Therefore, we employed RoBERTa to obtain language embedding. We used huggingface japanese-roberta-base pretrained model, which outputs a 768-dimensional vector for each word or token.

4.1.2 Facial features. We used OpenFace [3], which is a computer-vision-based toolkit capable of facial landmark detection, head-pose estimation, and eye-gaze estimation. It also outputs facial action units (AUs) [9], which were proposed in the Facial Action Coding System (FACS) [10] to encode the fundamental actions of individuals or groups of muscles typically observed while producing facial expressions. By applying OpenFace to the client and counselor videos, we extracted the following features and obtained 674 dimensional vectors: 2D and 3D facial landmarks, 2D and 3D gaze direction, 3D head poses (location and rotation), and intensities and occurrences of facial AUs. All the feature sets were z-normalized for the client and counselor.

Table 1: Sample Dialogue with MISC Code. FN: Follow/Neutral, CT: Change Talk, ST: Sustain Talk

Speaker	Category	Utterance
Client	FN	I eat instant food once a day
Client	FN	I like it, and, one day later
Client	FN	I can eat it. This is my rule, and I've been living like that for a long time.
Client	CT	So, I tend to reach for those things easily, and that's not good.
Client	ST	I'm already aware of that, but it's hard to stop.
Counselor	Reflect	Potato chips is a kind of vegetable in your life.
Client	FN	It's not an exaggeration to say that they're already a part of my life, and I've been eating them since I was a child.
Counselor	Facilitate	Um

Table 2: Client utterance distribution

Change Talk	Sustain Talk	Follow/Neutral
15%	9%	76%

4.2 Proposed Models

We propose CT detection multimodal models by fusing language and facial features described in Section 4.1. Following previous studies, we employed a recurrent neural network approach that classified client utterances into three classes (ST, CT, and FN). In this study, we merged ST and FN into one category, and created a 2-class classification model (CT and not-CT). By employing the long and short-term memory (LSTM) approach [13], we used a one-layer bidirectional LSTM (BiLSTM) to encode the language and facial features of the target and context utterances, and concatenated the encoder outputs as a multimodal representation. The network architecture of this model is illustrated in Figure 2.

4.2.1 Language + Client Facial information (BiLSTM_Lang+ClFace). To obtain the language embedding for each utterance, we used mean pooling: the average of the last hidden state of all tokens in the output of the RoBERTa Japanese pretrained model (Section 4.1.1). In addition to the target client utterance, we computed language embedding for five context utterances preceding the target utterance. We decided on the size of the context based on empirical analyses. The context utterances included counselor utterances. To indicate the speaker in the language embedding, we added a binary dimension at the beginning of each language embedding. Therefore, language embedding has 769 dimensions. Subsequently, the language vectors were input to BiLSTM to encode the target and context utterances as a sequence. The advantage of LSTM is that it maintains long-range connections between values in a sequence. The dimensions of the hidden state of the network were 300. The last hidden states produced from the bidirectional process were concatenated, and a 600-dim tensor was obtained as a language representation.

For facial information, as described in Section 4.1.2, a 674-dimensional vector was obtained from OpenFace for each frame of the client video. Such facial feature vectors were extracted for a given utterance by examining the time stamps of the start and end times of the utterance, and the average values were used as the facial embedding of the utterance. Similar to language embedding, facial embedding was created for the target client utterance and five context utterances. As this model used only client facial information, client facial embeddings were used for all context utterances, even if a context utterance was spoken by the counselor. Thus, in such cases, facial embeddings represent the client's facial information while listening to the counselor's utterance. A sequence of facial embeddings for the target and context utterances was then input into the BiLSTM to compute the facial representation. The dimensions of the hidden state of the network were 300. The last hidden states produced from the bidirectional process were concatenated, and a 600-dim tensor was obtained. This tensor was fed to the FC layer to reduce the tensor dimension to 200 dim. This facial representation was concatenated with the output of the language BiLSTM (600 dim). The 800 dimensional multimodal tensor was fed to an FC layer to downsize the vector to 300 dim. This 300-dim vector is applied to a classification layer, which is followed by a softmax layer to make 2-class classification (CT/not-CT) and calculate the cross-entropy loss. The reason for using a larger number of units for language representation than for facial representation is that we assumed that language information is more useful for classifying spoken utterances and that facial information complements language information.

4.2.2 Language + Counselor Facial information (BiLSTM_Lang+CoFace). This model uses the same network as the BiLSTM_Lang+ClFace model, as illustrated in Figure. 2. The only difference is that the counselor's facial information is used instead of client information. Thus, the facial information represents the counselor's facial expressions during a sequence of utterance exchanges, including the client's target utterance.

4.2.3 Language + Client and Counselor Facial information (BiLSTM_Lang+Face). This model uses both client and counselor facial information, each of which is applied to the BiLSTM encoder in the same manner as in the BiLSTM_Lang+ClFace and BiLSTM_Lang+CoFace models. Each encoder outputs a 600-dimensional tensor, and these two tensors are concatenated and fed to an FC layer. Consequently, a 300-dim vector was created as the facial information vector. This vector was then concatenated with the 600-dim language vector, which is the same as in the other models, and applied to an FC layer to produce a 300-dim vector. This vector was used in the classification layer.

Table 3: Distribution of the client answers to a question "When will you change your behavior?"

today	tomorrow	this week	this month	not sure	cannot change
9	16	10	6	5	2

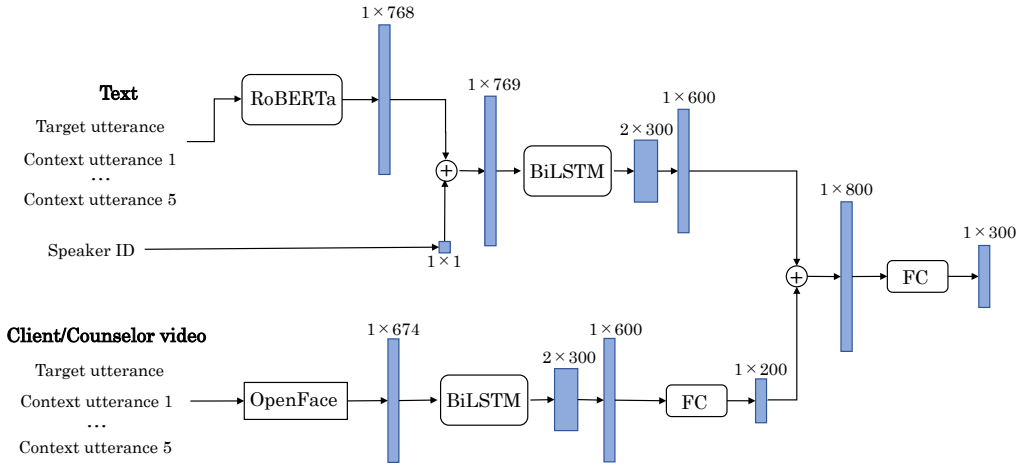

Figure 2: Network architecture for language and facial information multimodal models

4.3 Ablations

To conduct an ablation study, we created unimodal models, models without context, and models with a different architecture.

4.3.1 *Unimodal models.*

BiLSTM_Lang We created BiLSTM language unimodal model only using the language representation. The 600-dim tensor output from the BiLSTM language encoder was fed to a fully connected layer (FC) to reduce the tensor dimension to 300. This vector was applied to a binary classification layer.

BiGRU_Lang In order to compare with previous studies [25, 26], we created a language unimodal model using GRU, which has fewer gates than LSTM, by simply replacing the BiLSTM encoder in the BiLSTM_Lang model with BiGRU.

FC_Lang As RoBERTa itself is a powerful language model, we created a model without RNN-based encoder, where the language embedding of the target utterance obtained from RoBERTa was directly input to an FC layer.

BiLSTM_Face This model uses facial information for both participants (client and counselor) but does not use language information. In this model, the 300-dim facial information vector created in the BiLSTM_Lang+Face model was used in the classification layer without concatenating it with the language information.

4.3.2 *Multimodal models for comparison.*

BiGRU_Lang+Cl/CoFace These models were created by replacing BiLSTM with BiGRU in BiLSTM_Lang+ClFace and BiLSTM_Lang+CoFace models.

FC_Lang+ClFace This model does not use context utterances and RNN-based encoder. Language embedding and facial embedding for the target utterance were concatenated and then input to an FC layer.

5 EXPERIMENTS

We conducted experiments by creating a dataset for MI sessions using the corpus described in Section 3. We created 10 deep neural network (DNN) models by combining language and facial information from the client and counselor described in Section 4 and evaluated the performance.

5.1 Experimental setup

Our corpus included 48 MI sessions for 48 different clients. We used 39 sessions (81 %) for training, five (10 %) for validation, and four for testing (8 %). As the distribution was imbalanced, we calculated cross-entropy loss by setting class weight to 1.0 for FN and 4.0 for CT. The weights were determined based on the proportion of CT to FN. We used Adadelta to optimize the model and trained each model for 300 epochs. The batch size was 24. The best model was chosen based on the F1 score of CT in the validation set.

5.2 Results

The experimental results are shown in Table 4. As the distribution of the client utterance category was imbalanced, we evaluated our models with and without resampling the test set. In resampling the test set, to ensure that our dataset has a similar distribution to other MI datasets [25], we downsampled the FN utterances. The proportion of CT cases was approximately 0.28. We ran the test

Table 4: Model performance: Numbers in parenthesis indicate the results with resampled test set

	CT			FN			
	Precision	Recall	F1-score	Precision	Recall	F1-score	F1-macro
FC_Lang	0.463(0.586)	0.660(0.660)	0.544(0.621)	0.919(0.861)	0.834(0.818)	0.874(0.839)	0.709(0.730)
BiGRU_Lang	0.384(0.502)	0.739(0.739)	0.506(0.597)	0.929(0.876)	0.743(0.716)	0.826(0.788)	0.666(0.692)
BiLSTM_Lang	0.312(0.421)	0.987(0.987)	0.474(0.590)	0.995(0.990)	0.528(0.478)	0.690(0.645)	0.582(0.617)
BiLSTM_Face	0.168(0.272)	0.137(0.137)	0.151(0.182)	0.820(0.721)	0.852(0.858)	0.836(0.783)	0.493(0.483)
FC_Lang+ClFace	0.442(0.537)	0.647(0.647)	0.525(0.586)	0.915(0.868)	0.823(0.809)	0.866(0.838)	0.696(0.712)
BiGRU_Lang+ClFace	0.525(0.616)	0.542(0.542)	0.534(0.577)	0.9(0.832)	0.894(0.870)	0.897(0.850)	0.715(0.714)
BiGRU_Lang+CoFace	0.454(0.580)	0.582(0.582)	0.510(0.581)	0.903(0.841)	0.848(0.839)	0.875(0.840)	0.692(0.710)
BiLSTM_Lang+ClFace	0.475(0.607)	0.804(0.804)	**0.600(0.692)**	0.950(0.913)	0.807(0.799)	0.873(0.852)	**0.735(0.772)**
BiLSTM_Lang+CoFace	0.493(0.595)	0.699(0.699)	0.578(0.643)	0.928(0.876)	0.844(0.816)	0.884(0.845)	0.731(0.744)
BiLSTM_Lang+Face	0.452(0.566)	0.497(0.497)	0.474(0.529)	0.888(0.812)	0.870(0.851)	0.879(0.831)	0.676(0.680)

five times and calculated the average. When we tested our models without resampling, we used all samples in test set. The proportion of CT was 0.16. As our goal was to detect Change Talk in client utterances, the F1 score for CT was the most important metric. The BiLSTM_Lang+ClFace model achieves the highest F1 score (0.6 (0.692)). This model also achieves the best F1-macro (0.735 (0.772)). The second best model was the BiLSTM_Lang+CoFace model (F1 score for CT:0.578 (0.643)), F1-macro:0.731 (0.744)). The BiLSTM_Lang+Face model, which includes facial information for both participants, did not perform well compared to BiLSTM_Lang+ClFace and BiLSTM_Lang+CoFace.

Among the unimodal models, the FC_Lang model performed best (0.544(0.621)), suggesting that the RoBERTa language model produces useful embedding to detect CT, and encoding language information in context utterances using RNN-based network (BiLSTM_Lang and BiGRU_Lang) did not contribute to improving the model performance. Apparently, face unimodal model (BiLSTM_Face) was improperly trained, suggesting that facial information alone is insufficient for detecting CT, but is useful when combining language information.

In multimodal models for comparison, encoding facial features using BiGRU is not very useful. The F1 scores of the BiGRU multimodal models (BiGRU_Lang+ClFace:0.534(0.577) and BiGRU_Lang+CoFace:0.510(0.581)) are lower than those of the BiLSTM models. This suggests that BiLSTM is more efficient than BiGRU in encoding facial information over context utterances. As described above, BiLSTM_Lang+ClFace reached the highest F1-score for CT, followed by the BiLSTM_Lang+CoFace model. To examine whether these two proposed models significantly outperformed other comparison models, we conducted t-test using the resampled test set. Figure 3 shows the t-test results and boxplots that visualize the distribution, where the F1 score for the CT class in the best model (BiLSTM_Lang+ClFace) was significantly higher than that of the BiLSTM_Lang+CoFace model, which was the second best model ($t(8)=10.916$, $p<0.01$). Moreover, this BiLSTM_Lang+CoFace model performed significantly better than the FC_Lang model, which was the best among other models ($t(8)=3,8$), $p<0.01$). These results showed that fusing language information with client or counselor facial information in preceding utterances improves the model performance. It was also found that client facial information was more useful than counselor facial information.

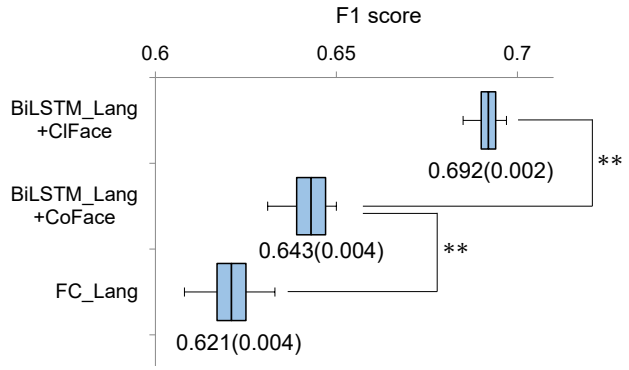

Figure 3: Results of t-test and distribution visualization: The center line of the boxplots indicates mean F1 score for CT. The upper and lower end of the box indicate one standard deviation above and below the mean, and the upper and lower whisker indicate the maximum and minimum values, respectively. ** indicates the difference is statistically significant at the 0.01 level ($p<0.01$) in t-test. The numbers in the graph indicate mean F1 score and the numbers in parenthesis are standard deviation.

5.3 Discussion

Previous studies on client utterance classification have tackled 3-class classification task: CT, ST, and FN. As reviewed in Section 2.2, Can et al. [4] employed CRF in client utterance classification as part of their MISC coding task: automatic coding for both counselor and client utterances. Xiao et al. [32] proposed a BiGRU model that uses a sequence of word embeddings for the target utterance as input and predicts the utterance category. More recently, Tavabi et al. [26] used RoBERTa embeddings and employed GRU to encode utterance history. Cao et al. [5] proposed BiGRU model with dialogue history and used word- and sentence-level attention.

Table 5 shows the model performance reported in these previous studies as well as the performance of our BiLSTM_Lang+ClFace model (without resampling). The table shows F1-macro and F1 scores for CT. It also shows the proportion of CT and FN as the measure of label imbalance because these studies used different

Table 5: Impact of label imbalance on performance

	F1 macro	F1 score for CT	CT proportion	FN proportion
Can et al.(2015) [4]	0.44	0.2	0.09	0.82
Xiao et al.(2016) [32]	0.5	0.3	0.10	0.80
Cao et al.(2019) [5]	0.54	0.39	0.09	0.83
Tavabi et al.(2021) [26]	0.66	0.64	0.28	0.59
BiLSTM_Lang+ClFace	0.74	0.6	0.16	0.74

Table 6: Sample of a client turn

Preidcted	Ground Truth	Utterance
CT	FN	Our original goal, we can save money.
CT	FN	I don't know how to call it.
CT	CT	I can set a goal.
CT	CT	I personally think it's an advantage to be able to have a strong will.

datasets. All these studies found a label imbalance problem, and model performance seems to depend not only on the model architecture but also on the severity of the label imbalance problem. As shown in the table, our label imbalance is less severe than that in the first three studies [4, 5, 32], but more severe than [26]. Although it may not be appreciated to compare the performance of our 2-class classification model with 3-class classification models of other studies, and we also used our new dataset, this table still clearly shows that the label imbalance problem impacts the F1 score of the CT category. It dropped to 0.6 when using all samples in the test set because of an increase of false positives.

As another evaluation method, we used a turn as the unit of analysis. When we analyzed the errors, it was found that some utterances were very short fragments. This is because we recognized an utterance boundary when a silent interval longer than 200 ms was detected. In such cases, a client turn consisted of a sequence of utterances and judging the category for every utterance is difficult. In addition, we found that in many cases, statements that communicated the client's intention were observed at the end of the turn, and the last utterance of the turn more clearly expressed CT. Table 6 shows such an example. In the first two utterances, the ground truth was FN but changed to CT for the last two utterances in the turn. Our model recognized CT slightly earlier than the ground truth.

Based on this observation, we combined a sequence of client utterances as a turn. When the speaker changed, it was the end of the previous turn (even if the next counselor utterance was a short acknowledgement, it was regarded as a turn change). We then assigned a label to each turn using the following rules: (1) when, in the ground truth, at least one CT is included in a turn, CT label is assigned as the true value of that turn, (2) if the predicted sequence ends with CT, CT is assigned as the predicted label for that turn. (3) if there is a matched CT sequence longer than five utterances, such case is counted as a true positive. The example in Table 6 was counted as a true-positive case by applying rules (1) and (2).

As a result of evaluating the model using these rules, the F1 score of CT category was 0.65, which is slightly better than 0.64 reported in a previous study (Table 5). A more important contribution of this study is that we found that facial information, especially for clients, is useful in detecting CT, and the frequency of CT is a good predictor of the client's motivation level immediately after the session, as discussed in Section 3.4.3.

6 CONCLUSIONS

With a goal of assisting MI in nutrition and fitness domains, this study proposed a BiLSTM model that detects client CT and addressed a question whether facial information contributes to this task. For this purpose, we first collected MI encounters in nutrition and fitness domains, and created a corpus with MISC code annotation. We also found that clients were more motivated when they uttered more CT during the MI encounter, suggesting that detecting CT is useful in estimating the client's motivation level. We then proposed multimodal neural network models that detect CT. The experimental results showed that fusing language and facial information improved model performance. In particular, client facial information improved model performance most effectively.

6.1 Limitations

In our experiment, BiLSTM_Lang+ClFace and BiLSTM_Lang+CoFace were the two best models for CT detection. However, BiLSTM_ Lang+Face, which uses facial information from both the client and counselor, did not perform well. It is possible that our DNN architecture did not successfully model the facial signals from both participants. In addition, language unimodal models in previous studies [5, 11] used attention mechanism; however, our model did not exploit this mechanism in modeling the sequence of utterances. Improving the representation of multimodal context by employing multimodal transformer and cross-modal attention [28, 35] would be a future direction.

6.2 Future Work

As a future direction, a detailed analysis of facial information is necessary to investigate whether there are any typical facial expressions when clients communicate CT. Another next step is a more in-depth analysis of the content of CT. We did not use subcategories of CT, such as Reason+, Ability+. A more detailed analysis is expected to get better understanding of client status and attitude, which will be useful for therapists. It is also necessary to improve model performance. Label imbalance is one of the main issues. Therefore, a more sophisticated sampling method that covers a variety of samples with different characteristics would be necessary.

ACKNOWLEDGMENTS

This work was supported by JST Moonshot R&D Grant Number JPMJMS2011 and JST AIP Trilateral AI Research (PANORAMA project, grant no. JPMJCR20G6) and JSPS KAKENHI (grant numbers JP19H01120 and JP19H04159).

REFERENCES

[1] Chanuwas Aswamenakul, Joshua Woolley, Lixing Liu, Stefan Scherer, Kate B. Carey, and Brian Borsari. 2018. Multimodal analysis of client behavioral change coding in motivational interviewing. *ICMI 2018 - Proceedings of the 2018 International Conference on Multimodal Interaction* (2018), 356–360. https://doi.org/10.1145/3242969.3242990

[2] Tadas Baltru?aitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 41, 2 (Feb 2019), 423–443.

[3] Tadas Baltrušaitis, Amir Zadeh, Yao Lim Chong, and Louis-Philippe Morency. 2018. OpenFace 2.0 : Facial Behavior Analysis Toolkit. In *IEEE International Conference on Automatic Face and Gesture Recognition*.

[4] Dogan Can, David C. Atkins, and Shrikanth S. Narayanan. 2015. A dialog act tagging approach to behavioral coding: A case study of addiction counseling conversations. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2015-January (2015), 339–343. https://doi.org/10.21437/interspeech.2015-151

[5] Jie Cao, Michael Tanana, Zac E. Imel, Eric Poitras, David C. Atkins, and Vivek Srikumar. 2019. Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 5599–5611.

[6] Dawn Clifford and L. Curtis. 2016. *Motivational Interviewing in Nutrition and Fitness*. Guilford Publications.

[7] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert Rizzo, and Louis-Philippe Morency. 2014. SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support. In *Proc. International Conference on Autonomous Agents and Multi-agent Systems*. 1061–1068.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[9] Shichuan Du, Yong Tao, and Aleix M. Martinez. 2014. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences of the United States of America* 111, 15 (2014). https://doi.org/10.1073/pnas.1322355111

[10] Paul Ekman and Wallace V. Friesen. 1978. Facial Action Coding System: A Technique for the Measurement of Facial Movement.

[11] James Gibson, Doğan Can, Panayiotis Georgiou, David C. Atkins, and Shrikanth Narayanan. 2017. Attention networks for modeling behaviors in addiction counseling. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2017-August, August (2017), 3251–3255. https://doi.org/10.21437/Interspeech.2017-218

[12] K. M. Griffin and M. A. Sayette. 2008. Facial reactions to smoking cues relate to ambivalence about smoking. *Psychology of addictive behaviors : journal of the Society of Psychologists in Addictive Behaviors* 22, 4 (2008), 551–556.

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (nov 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[14] Christine L. Lisetti, Ugan Yasavur, Ubbo Visser, and Naphtali Rishe. 2011. Toward conducting motivational interviewing with an on-demand clinician avatar for tailored health behavior change interventions. In *2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*. 246–249. https://doi.org/10.4108/icst.pervasivehealth.2011.246078

[15] 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. https://doi.org/10.48550/ARXIV.1907.11692

[16] S. Michie, R. West, R. Campbell, J. Brown, and H. Gainforth. 2014. *ABC of Behaviour Change Theories*. Silverback Publishing. https://books.google.fr/books?id=WQ7SoAEACAAJ

[17] William R. Miller, Theresa B. Moyers, Denise Ernst, and Paul Amrhein. 2003. Manual for the motivational interviewing skill code (MISC).

[18] W. R. Miller and Rollnick S. 2013. *Motivational Interviewing. Helping people change*. Guilford Press.

[19] Chathurika Palliya Guruge, Sharon Oviatt, Pari Delir Haghighi, and Elizabeth Pritchard. 2021. Advances in multimodal behavioral analytics for early dementia diagnosis: A review. 328–340.

[20] Anastasia Pampouchidou, Panagiotis G. Simos, Kostas Marias, Fabrice Meriaudeau, Fan Yang, Matthew Pediaditis, and Manolis Tsiknakis. 2019. Automatic Assessment of Depression Based on Visual Cues: A Systematic Review. *IEEE Transactions on Affective Computing* 10, 4 (2019), 445–470. https://doi.org/10.1109/TAFFC.2017.2724035

[21] Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence An, Kathy J. Goggin, and Delwyn Catley. 2017. Predicting counselor behaviors in motivational interviewing encounters. *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference* 2 (2017), 1128–1137. https://doi.org/10.18653/v1/e17-1106

[22] Giota Stratou and Louis-Philippe Morency. 2017. MultiSense—Context-Aware Nonverbal Behavior Analysis Framework: A Psychological Distress Use Case. *IEEE Trans. on Affective Computing* 8, 2 (2017), 190–203.

[23] Michael Tanana, Kevin Hallgren, Zac Imel, David Atkins, Padhraic Smyth, and Vivek Srikumar. 2015. Recursive Neural Networks for Coding Therapist and Patient Behavior in Motivational Interviewing. *2nd Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych 2015 - Proceedings of the Workshop* (2015), 71–79. https://doi.org/10.3115/v1/w15-1209

[24] Leili Tavabi, Kalin Stefanov, Setareh Nasihati Gilani, David Traum, and Mohammad Soleymani. 2019. Multimodal Learning for Identifying Opportunities for Empathetic Responses. In *Proc. ACM International Conference on Multimodal Interaction*. 95–104.

[25] Leili Tavabi, Kalin Stefanov, Larry Zhang, Brian Borsari, Joshua D. Woolley, Stefan Scherer, and Mohammad Soleymani. 2020. *Multimodal Automatic Coding of Client Behavior in Motivational Interviewing*. Association for Computing Machinery, New York, NY, USA, 406?413. https://doi.org/10.1145/3382507.3418853

[26] Leili Tavabi, Trang Tran, Kalin Stefanov, Brian Borsari, Joshua D. Woolley, Stefan Scherer, and Mohammad Soleymani. 2021. Analysis of Behavior Classification in Motivational Interviewing. *Computational Linguistics and Clinical Psychology: Improving Access, CLPsych 2021 - Proceedings of the 7th Workshop, in conjunction with NAACL 2021* (2021), 110–115. https://doi.org/10.18653/v1/2021.clpsych-1.13

[27] Moyers TB, Martin T, Houck JM, Christopher PJ, and Tonigan JS. 2009. From in-session behaviors to drinking outcomes: a causal chain for motivational interviewing. *J Consult Clin Psychol.* 77, 6 (Dec 2009), 1113–24. https://doi.org/10.1037/a0017189

[28] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 6558–6569. https://doi.org/10.18653/v1/P19-1656

[29] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and vision computing* 27, 12 (2009), 1743–1759.

[30] Philipp Werner, Daniel Lopez-Martinez, Steffen Walter, Ayoub Al-Hamadi, Sascha Gruss, and Rosalind W. Picard. 2022. Automatic Recognition Methods Supporting Pain Assessment: A Survey. *IEEE Transactions on Affective Computing* 13, 1 (2022), 530–552. https://doi.org/10.1109/TAFFC.2019.2946774

[31] Zixiu Wu, Rim Helaoui, Vivek Kumar, Diego Reforgiato Recupero, and Daniele Riboni. 2020. Towards Detecting Need for Empathetic Response in Motivational Interviewing. In *Companion Publication of the International Conference on Multimodal Interaction* (Virtual Event, Netherlands). Association for Computing Machinery, New York, NY, USA, 497?502. https://doi.org/10.1145/3395035.3425228

[32] Bo Xiao, Doğan Can, James Gibson, Zac E. Imel, David C. Atkins, Panayiotis Georgiou, and Shrikanth Narayanan. 2016. Behavioral coding of therapist language in addiction counseling using recurrent neural networks. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 08-12-September-2016, September (2016), 908–912. https://doi.org/10.21437/interspeech.2016-1560

[33] Bo Xiao, Zac E Imel, Panayiotis Georgiou, David C Atkins, and Shrikanth S Narayanan. 2016. Computational analysis and simulation of empathic behaviors: A survey of empathy modeling with behavioral signal processing framework. *Current psychiatry reports* 18, 5 (2016), 1–11.

[34] Bo Xiao, Zac E Imel, Panayiotis G Georgiou, David C Atkins, and Shrikanth S Narayanan. 2015. "Rate My Therapist": Automated Detection of Empathy in Drug and Alcohol Counseling via Speech and Language Processing. *PLOS ONE* 10, 12 (2015), 1–15.

[35] Kaicheng Yang, Hua Xu, and Kai Gao. 2020. CM-BERT: Cross-Modal BERT for Text-Audio Sentiment Analysis. *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia* (2020), 521–528. https://doi.org/10.1145/3394171.3413690