Analysis of Behavior Classification in Motivational Interviewing

Leili Tavabi[1], **Trang Tran**[1], **Kalin Stefanov**[2],
Brian Borsari[3], **Joshua D Woolley** [3], **Stefan Scherer**[1], **Mohammad Soleymani**[1]
[1]Institute for Creative Technologies, University of Southern California, Los Angeles, CA, USA
[2]Faculty of Information Technology, Monash University, Melbourne, VIC, Australia
[3]VA Hospital San Francisco, University of California San Francisco, San Francisco, CA, USA
{ltavabi,ttran}@ict.usc.edu, kalin.stefanov@monash.edu,
Brian.Borsari@va.gov, josh.woolley@ucsf.edu,
{scherer,soleymani}@ict.usc.edu

Abstract

Analysis of client and therapist behavior in counseling sessions can provide helpful insights for assessing the quality of the session and consequently, the client's behavioral outcome. In this paper, we study the automatic classification of standardized behavior codes (i.e. annotations) used for assessment of psychotherapy sessions in Motivational Interviewing (MI). We develop models and examine the classification of client behaviors throughout MI sessions, comparing the performance by models trained on large pretrained embeddings (RoBERTa) versus interpretable and expert-selected features (LIWC). Our best performing model using the pretrained RoBERTa embeddings beats the baseline model, achieving an F1 score of 0.66 in the subject-independent 3-class classification. Through statistical analysis on the classification results, we identify prominent LIWC features that may not have been captured by the model using pretrained embeddings. Although classification using LIWC features underperforms RoBERTa, our findings motivate the future direction of incorporating auxiliary tasks in the classification of MI codes.

1 Introduction

Motivational Interviewing (MI) is a psychotherapy treatment style for resolving ambivalence toward a problem such as alcohol or substance abuse. MI approaches focus on eliciting clients' own intrinsic reasons for changing their behavior toward the desired outcome. MI commonly leverages a behavioral coding (annotation) system, Motivational Interviewing Skills Code (MISC) (Miller et al., 2003), which human annotators follow for coding both client's and therapist's utterance-level intentions and behaviors. These codes have shown to be effective means of assessing the quality of the session, training therapists, and estimating clients' behavioral outcomes (Lundahl et al., 2010; Diclemente et al., 2017; Magill et al., 2018). Due to the high cost and labor-intensive procedure of manually annotating utterance-level behaviors, existing efforts have worked on automatic coding of the MI behaviors. The client utterances throughout the MI session are categorized based on their expressed attitude toward change of behavior: (1) Change Talk (CT): willing to change, (2) Sustain Talk (ST): resisting to change, and (3) Follow/Neutral (FN): other talk unrelated to change. An example conversation between a therapist (T) and a client (C) is shown below.

- T: [...] you talked about drinking about 7 times a week [...] Does that sound about right, or?
- C: I don't know so much any, like 5, probably like, the most 4 now, in the middle of the week I try to just kinda do work, **(CT)**
- C: I mean, like I would **(ST)**
- C: but, but getting up's worse, it's like being tired, not so much hungover just feeling uhh, class. [...] **(CT)**
- T: When you do drink, how much would you say, would you say the ten's about accurate?
- C: About around ten, maybe less, maybe more, depends like, I don't really count or anything but, it's probably around ten or so. **(FN)**

Previous work in MI literature mainly approached automatic classification of behavior codes in MI by modeling utterance-level representations. Aswamenakul et al. (2018) trained a logistic regression model using both interpretable linguistic features (LIWC) and GloVe embeddings, finding that Sustain Talk is associated with positive attitude towards drinking, and the opposite for Change Talk. To account for dialog context, Can et al. (2015) formulated the task as a sequence labeling problem, and trained a Conditional Random Field (CRF) to predict MI codes. More recent approaches leveraged advances in neural networks, using standard recurrent neural networks (RNNs) (Xiao et al., 2016; Ewbank et al., 2020; Gibson

Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology, pages 110–115

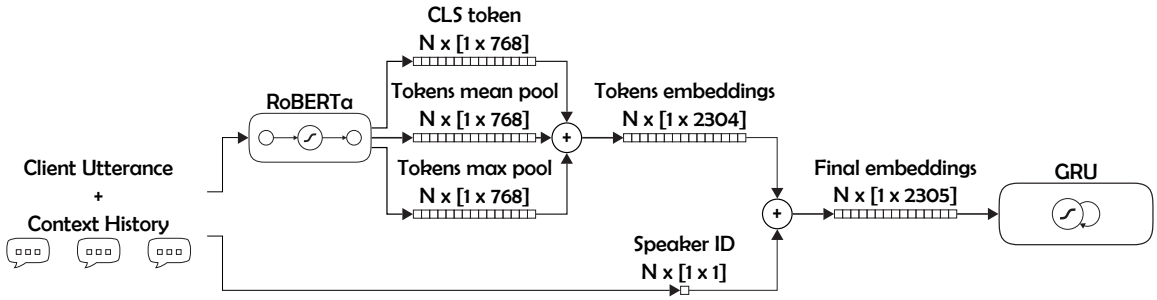

Figure 1: Utterance representation from RoBERTa embeddings.

et al., 2016; Huang et al., 2018) or hierarchical encoders with attention (Cao et al., 2019). In addition to context modeling, Tavabi et al. (2020) leveraged pretrained contextualized embeddings (Devlin et al., 2019) and incorporated the speech modality to classify MI codes, beating the previous baseline of Aswamenakul et al. (2018) on a similar dataset. The most gain seemed to come from powerful pretrained embeddings, as with many other NLP tasks. However, it is unclear what these BERT-like embeddings learn, as they are not as interpretable as the psycholinguistically motivated features (LIWC).

In this paper, we study the quality of automatic MI coding models in an attempt to understand what distinguishes language patterns in Change Talk, Sustain Talk, and Follow/Neutral. We develop a system for classifying clients' utterance-level MI codes by modeling the client's utterance and the preceding context history from both the client and the therapist. We compare the effectiveness and interpretability between contextualized pretrained embeddings and hand-crafted features, by training classifiers using (1) pretrained RoBERTa embeddings (Liu et al., 2019), (2) an interpretable and dictionary-based feature set, Linguistic Inquiry Word Count (LIWC) (Pennebaker et al., 2001). Our best-performing model outperforms the baseline model from previous work on the same dataset (Tavabi et al., 2020), reaching F1=0.66 from F1=0.63.

In examining misclassifications by both models, we identify features that are significant across classes. Our findings suggest that large pretrained embeddings like RoBERTa, despite their high representation power, might not necessarily capture all the salient features that are important in distinguishing the classes. We identified prominent features that are statistically significant across classes on the entire dataset, as well as the misclassified samples. Theses findings suggest that our systems might benefit from fine-tuning pretrained embeddings, adding auxiliary tasks (e.g sentiment classification), and better context modeling.

2 Data

We use two clinical datasets (Borsari et al., 2015) collected in college campuses from real MI sessions with students having alcohol-related problems. The data consists of transcripts and audio recordings from the client-therapist in-session dialogues. The sessions are manually transcribed, and labelled per utterance using MISC codes. The dataset includes 219 sessions for 219 clients, consisting of about 93k client and therapist utterances; the client-therapist distribution of utterances is 0.44-0.54. The dataset is highly imbalanced, with a class distribution of [0.13, 0.59, 0.28] for [Sustain Talk, Follow/Neutral, Change Talk]. In addition to the in-session text and speech data, the dataset consists of session-level measures regarding clients' behavioral changes toward the desired outcome. Additional metadata includes session-level global metrics such as therapist empathy, MI spirit, and client engagement.

3 Methodology

3.1 Embeddings and Feature sets

Pretrained RoBERTa Embeddings. RoBERTa (Liu et al., 2019) is an improved representation based on BERT (Devlin et al., 2019). RoBERTa differs from BERT in several aspects: removal of the Next Sentence Prediction objective, introduction of dynamic masking, pretrained on a larger dataset with larger mini-batches and longer sequences. These changes can improve the representations on our data, especially since dialogue utterances in psychotherapy can consist of very long sequences. Our preliminary experiments for fine-tuning both BERT and RoBERTa on our task showed that RoBERTa performed better. We therefore select

RoBERTa to obtain utterance representations.

Interpretable LIWC Features. LIWC (Pennebaker et al., 2001) is a dictionary-based tool that assigns scores in psychologically meaningful categories including social and affective processes, based on words in a text input. It was developed by experts in social psychology and linguistics, and provides a mechanism for gaining interpretable and explainable insights in the text input. Given our focus domain of clinical psychology, where domain knowledge is highly valuable, we select the psychologically-motivated LIWC feature set as a natural point of comparison.

3.2 Classification Model

For classifying the clients' MI codes, we learn the client utterance representation using features described in 3.1, as well as the preceding history from both the client and therapist. The input window includes the current utterance, and history context. Specifically, the input window consists of a total of 3 or more turn changes across speakers, where each turn consists of one or more consecutive utterances per speaker. In the beginning of the session, where the history context is shorter than the specified threshold, the context history consists of those limited preceding utterances. The size of the context window was selected empirically among 3, 4 or 5 turn changes.

Our input samples contain between 6 and 28 utterances depending on the dynamic of the dialogue, e.g. an example input could be [T C T T T C C T C], where T denotes Therapist's utterance and C denotes Client's. The motivation for using the entire window of context and final utterance is that the encoding by our recurrent neural network (RNN) would carry more information from the final utterance and closer context, while retaining relevant information from the beginning of the window. We also investigated encoding the current utterance separate from the context using a linear layer, but did not see improvements in the classification results.

For RoBERTa embeddings, each utterance representation is the concatenation of (1) CLS token (2) mean pooling of the tokens from the last hidden state (3) max pooling of the tokens from the last hidden state. Figure 1 illustrates this process. For LIWC representations, the features are already extracted on the utterance level. Additionally, for both RoBERTa and LIWC representations, we add a binary dimension for each utterance to indicate the speaker. The history context representation for both RoBERTa and LIWC is obtained by concatenating the utterance-level representation vectors into a 2d matrix. These inputs are then fed into a unidirectional GRU, and the last hidden state is used for the last classification layer.

4 Results and Discussions

For training, we use a 5-fold subject-independent cross validation. 10% of the train data from each fold is randomly selected in stratified fashion, and held out as the validation set. We optimize the network using AdamW (Loshchilov and Hutter, 2019), with a learning rate of 10^{-4} and batch size of 32. We train our model for 25 epochs with early stopping after 10 epochs, and select the model with the highest macro F1 on the validation set. To handle class imbalance, we use a cross-entropy loss with a weight vector inversely proportional to the number of samples in each class. The GRU hidden dimension is 256 and 32 when running on RoBERTa and LIWC representations, respectively.

We compare our work to the best performing model from previous work (Tavabi et al., 2020), trained on the same dataset and under the same evaluation protocol. Briefly, this baseline model differs from our current model in several aspects: BERT embeddings were used as input; the representation vector for the current client utterance is fed into a linear layer. The client and therapist utterances within the context window are separated, mean-pooled and fed individually to two different linear layers. The output encodings from the three linear layers are merged and fed into another linear layer before being passed to the classification layer.

We perform statistical analysis to identify prominent LIWC features across pairs of classes, as well as misclassified samples from each classifier. Since the classifiers encode context, we incorporate the context in the statistical analysis by averaging the feature vectors along utterances within the input window.

4.1 Classifier Performance

The classification results are shown in Table 1. The model trained using RoBERTa outperforms the model trained on LIWC features, in addition to beating the baseline model in (Tavabi et al., 2020) with F1-macro=0.66. Improved results over the

baseline model are likely due to the following: 1) The previous linear model encodes the client and therapist utterances from the context history separately, therefore potentially missing information from the dyadic interaction. 2) The RNN in our current model temporally encodes the dyadic interaction window. 3) Using RoBERTa embeddings improved over BERT embeddings, as RoBERTa was trained on larger datasets and on longer sequences, making them more powerful representations.

	Features		Baseline
	LIWC	RoBERTa	
ST	0.41	0.50	0.46
FN	0.78	0.84	0.81
CT	0.56	0.64	0.63
All (macro)	0.58	**0.66**	0.63
All (micro)	0.65	**0.74**	0.71

Table 1: F1-Score Classification Results

The results from other work on classifying client codes in MI range from F1-macro=0.44 (Can et al., 2015) to F1-macro=0.54 (Cao et al., 2019) on different datasets. Aswamenakul et al. (2018), who used a similar dataset to our work, reached F1-macro=0.57. Huang et al. (2018) obtained F1-macro=0.70 by using (ground truth) labels from prior utterances as the model input and domain adaptation for theme shifts throughout the session.

The F1 scores show that Sustain Talk, the minority class, is consistently the hardest to classify and Follow/Neutral, the majority class, the easiest. This is similar to findings from previous work in literature, e.g. (Can et al., 2015) and remains a challenge in automated MI coding. Using approaches like upsampling toward a more balanced dataset will be part of our future work. In order for these systems to be deployable in the clinical setting, the standard we adhere to is guided by a range developed by biostaticians in the field, which indicates values higher than 0.75 to be "excellent" (Cicchetti, 1994). Therefore, despite the good results, there is much room for improvement before such systems can be autonomously utilized in real-world MI sessions.

4.2 Error Analysis

Figure 2 shows the confusion matrices from classification results by the model using LIWC features vs. RoBERTa embeddings. Comparing between classes, Sustain Talk gets misclassified about equally as Follow/Neutral and Change Talk by RoBERTa but it is much more often misclassified as Change Talk by LIWC. On the other hand, Change Talk is more often misclassified as Follow/Neutral by RoBERTa, but misclassified as Sustain Talk by LIWC.

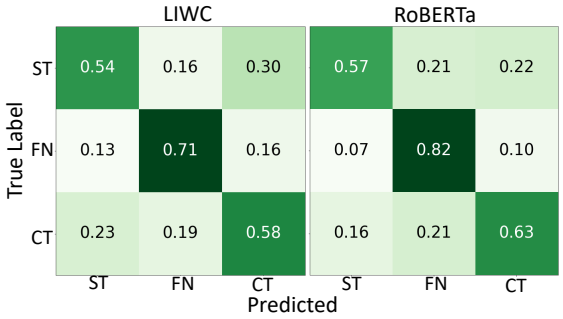

Figure 2: Confusion matrices (normalized by true labels) of classification results by LIWC (left) and RoBERTa (right) features.

Of the wrongly classified utterances by LIWC, 47% were correctly classified by RoBERTa. Of the RoBERTa misclassifications (11k utterances), about 30% were correctly classified by LIWC. Some examples of these cases are presented in Figure 3, which seem to be associated with certain key words related to salient features (Section 4.3).

T: What varies your drinking?
C: Money, **(CT → ST)**
C: if I have work to do I won't drink. **(CT)**
T: Okay.
...: ...
C: Anxious thing is kinda like I don't have control, like I, I'm shaky and stuff like that. **(CT)**
T: Ok. Is your heart racing faster or, and, and that type of thing?
C: No, it's not really anxious, it's kinda just like a ... **(CT → ST)**
T: It's more shaky?
C: It's like agitated, kind of. **(CT → ST)**

Figure 3: Example dialog with correct and incorrect classifications. T=therapist; C=client; red (true → predicted) denotes misclassification by RoBERTa but correctly classified by LIWC; blue (true label) denotes correct classification by both models.

When both RoBERTa and LIWC misclassified, they give the same wrong prediction on 70% of those utterances. Some anecdotal examples of such cases are shown in Figure 4, most seem to be highly context-dependent, suggesting that better modeling

of context would potentially be useful.

T: Oh, ok, so the summer you usually drink a little more
C: Yeah. **(FN)**
T: and then when you get to school, it's...
C: Kinda cut down a little bit. **(CT)**
T: I see, because of like, school and classes and stuff.
C: Yeah. **(CT → FN)**
T: And working on the weekends.
C: Yeah. **(CT → FN)**

Figure 4: Example dialog with correct and incorrect classifications. T=therapist; C=client. blue (true label) denotes correct classification by our models, red (true → predicted) denotes misclassification by both models.

We also experimented with simple concatenation of RoBERTa and LIWC features, but did not find significant improvements over the RoBERTa-only model. Better models for combining RoBERTa and LIWC features might improve our results, which will be part of future work.

4.3 Salient Features

Statistical analysis on LIWC features across the classes can help identify the salient features distinguishing the classes, therefore can signal important information picked up by the LIWC classifier. We used hierarchical Analysis of Variance (ANOVA), with talk types nested under sessions to account for individual differences, to find linguistic features that are significantly different across MI codes. To further examine the statistical significance across pairs of classes, we performed a Tukey post hoc test. We found the following features to be the most statistically different features across all the pairs of classes: 'WPS' (mean words per sentence), 'informal', 'assent' (e.g. agree, ok, yes), 'analytic.' Additionally, 'AllPunc' (use of punctuations) and 'function' (use of pronouns) were prominent features that were significantly distinguishing Follow/Neutral from the other classes.

We further looked into samples where RoBERTa representations might be limited (i.e. misclassified), while LIWC features were correct in the classification. Using ANOVA, we found the most prominent features in such samples across the 3 classes: 'swear' (6.06), 'money' (5.29), 'anger' (2.24), 'death' (2.19), and 'affiliation' (2.00), where numbers in parentheses denote F-statistic from hierarchical ANOVA. This is consistent with our error analysis in Section 4.2, as shown in Figure 3. The mean scores of the 'swear,' 'money,' and 'anger' categories are higher for Change Talk compared to other classes. We hypothesize that 'swear' and 'anger' in Change Talk may represent anger toward oneself regarding drinking behavior. Words in the 'money' category might be related to the high cost of alcohol (especially with college-age clients), which can be motivation for behavior change. The Change Talk samples misclassified by the RoBERTa model may indicate the model's failure to capture such patterns.

5 Conclusion

We developed models for the classification of clients' MI codes. We experimented with pre-trained RoBERTa embeddings and interpretable LIWC features as our model inputs, where the RoBERTa model outperformed the baseline from previous work, reaching F1=0.66. Through statistical analysis, we investigated prominent LIWC features that are significantly different across pairs of classes. We further looked into misclassified samples across the classifiers, and identified prominent features that may have not been captured by the RoBERTa model. This finding motivates the use of auxiliary tasks like sentiment and affect prediction, in addition to fine-tuning the model with domain-specific data and better context modeling.

With this work, we aim to develop systems for enhancing effective communication in MI, which can potentially generalize to other types of therapy approaches. Identifying patterns of change language can lead to MI strategies that will assist clinicians with treatment, while facilitating efficient means for training new therapists. These steps contribute to the long-term goal of providing cost- and time- effective evaluation of treatment fidelity, education of new therapists, and ultimately broadening access to lower-cost clinical resources for the general population.

Acknowledgments

This work was supported by NIAAA grants R01 AA027225, R01 AA017427 and R01 AA12518. The content is the responsibility of the authors and does not necessarily represent the official views of the NIAAA, NIH, Dept. of Veterans Affairs, or the US Government. We thank the clients and therapists for their audiotapes to be used in this work, and the anonymous reviewers for their feedback.

References

Chanuwas Aswamenakul, Lixing Liu, Kate B Carey, Joshua Woolley, Stefan Scherer, and Brian Borsari. 2018. Multimodal Analysis of Client Behavioral Change Coding in Motivational Interviewing. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 356–360.

Brian Borsari, Timothy R Apodaca, Kristina M Jackson, Nadine R Mastroleo, Molly Magill, Nancy P Barnett, and Kate B Carey. 2015. In-session processes of brief motivational interventions in two trials with mandated college students. *Journal of consulting and clinical psychology* 83, 1 (2015), 56.

Doğan Can, David C Atkins, and Shrikanth S Narayanan. 2015. A dialog act tagging approach to behavioral coding: A case study of addiction counseling conversations. In *Sixteenth Annual Conference of the International Speech Communication Association*.

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. *arXiv preprint arXiv:1907.00326* (2019).

Domenic V Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment* 6, 4 (1994), 284.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. NAACL*. 4171–4186.

C. Diclemente, Catherine M Corno, Meagan M. Graydon, Alicia E Wiprovnick, and Daniel J. Knoblach. 2017. Motivational Interviewing, Enhancement, and Brief Interventions Over the Last Decade: A Review of Reviews of Efficacy and Effectiveness. *Psychology of Addictive Behaviors* 31, 862–887.

MP Ewbank, R Cummins, V Tablan, A Catarino, S Buchholz, and AD Blackwell. 2020. Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts. *Psychotherapy Research* (2020), 1–13.

James Gibson, Dogan Can, Bo Xiao, Zac E Imel, David C Atkins, Panayiotis Georgiou, and Shrikanth Narayanan. 2016. A deep learning approach to modeling empathy in addiction counseling. *Commitment* 111 (2016), 21.

Xiaolei Huang, Lixing Liu, Kate Carey, Joshua Woolley, Stefan Scherer, and Brian Borsari. 2018. Modeling temporality of human intentions by domain adaptation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 696–701.

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. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Bkg6RiCqY7

Brad W Lundahl, Chelsea Kunz, Cynthia Brownell, Derrik Tollefson, and Brian L Burke. 2010. A meta-analysis of motivational interviewing: Twenty-five years of empirical studies. *Research on social work practice* 20, 2 (2010), 137–160.

Molly Magill, Timothy R Apodaca, Brian Borsari, Jacques Gaume, Ariel Hoadley, Rebecca EF Gordon, J Scott Tonigan, and Theresa Moyers. 2018. A meta-analysis of motivational interviewing process: Technical, relational, and conditional process models of change. *Journal of consulting and clinical psychology* 86, 2 (2018), 140.

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* (2003).

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

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. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 406–413.

Bo Xiao, Dogan Can, James Gibson, Zac E Imel, David C Atkins, Panayiotis G Georgiou, and Shrikanth S Narayanan. 2016. Behavioral Coding of Therapist Language in Addiction Counseling Using Recurrent Neural Networks.. In *Interspeech*. 908–912.