# Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes (Supplementary Materials)

**Jie Cao**[†], **Michael Tanana**[‡], **Zac E. Imel**[‡], **Eric Poitras**[‡], **David C. Atkins**[◇], **Vivek Srikumar**[†]
[†]School of Computing, University of Utah
[‡]Department of Educational Psychology, University of Utah
[◇]Department of Psychiatry and Public Health, University of Washington
{jcao, svivek}@cs.utah.edu,
{michael.tanana, zac.imel, eric.poitras}@utah.edu,
datkins@u.washington.edu

## A Appendix

**Different Clustering Strategies for MISC** The original MISC description of Miller et al. (2003) included 28 labels (9 client, 19 therapist). Due to data scarcity and label confusion, some labels were merged into a coarser set. Can et al. (2015) retain 6 original labels FA, GI, QUC, QUO, REC, RES, and merge remaining 13 rare labels into a single COU label, they merge all 9 client codes into a single CLI label. Instead, Tanana et al. (2016) merge only 8 of rare labels into a OTHER label and they cluster client codes according to the valence of changing, sustaining or being neutral on the addictive behavior(Atkins et al., 2014). Then Xiao et al. (2016) combine and improve above two clustering strategies by splitting the all 13 rare labels according to whether the code represents MI-adherent(MIA) and MI-nonadherent (MIN) We show more details about the original labels in MIA and MIN in Table 1

**Model Setup** We use 300-dimensional Glove embeddings pre-trained on 840B tokens from Common Crawl (Pennington et al., 2014). We do not update the embedding during training. Tokens not covered by Glove are using a randomly initialized UNK embedding. We also use character-level deep contextualized embedding ELMo 5.5B model by concatenating the corresponding ELMo word encoding after the word embedding vector. For speaker information, we randomly initialize them with 8 dimensional vectors and update them during training. We used a dropout rate of 0.3 for the embedding layers.

We trained all models using Adam (Kingma and Ba, 2015) with learning rate chosen by cross validation between $[1e^{-4}, 5 * 1e^{-4}]$, gradient norms clipping from at $[1.0, 5.0]$, and minibatch sizes of 32 or 64. We use the same hidden size for both utterance encoder, dialogue encoder and other attention memory hidden size; it has been selected from $\{64, 128, 256, 512\}$. We set a smaller dropout 0.2 for the final two fully connected layers. All the models are trained for 100 epochs with early-stoping based on macro $F_1$ over development results.

**Detailed Results of Our Main Models** In the main text, we only show the $F_1$ score of each our proposed models. We summarize the performance of our best models for both categorzing and forecasting MISC codes in Table 2 with precision, recall and $F_1$ for each codes.

**Domain Specific Glove and ELMo** We use the general psychotherapy corpus with 6.5M words (Alexander Street Press) to train the domain specific word embeddings **Glove**$_{psyc}$ with 50, 100, 300 dimension. Also, we trained ELMo with 1 highway connection and 256-dimensional output size to get **ELMo**$_{psyc}$. We found that ELMo 5.5B performs better than ELMo psyc in our experiments, and general Glove-300 is better than the **Glove**$_{psyc}$. Hence for main results of our models, we use **ELMo**$_{generic}$ by default. Please see more details in Table 3

**Full Results for Ablation on Forecasting Tasks** In addition to the ablation table in the main paper for categorizing tasks, we reported more ablation details on forecasting task in Table 4. Word-level attention shows no help for both client and therapist codes. While sentence-level attention helps more on therapist codes than on client codes. Multi-head self attention alsoachieves better performance than anchor-based attention in forecasting tasks.

**Label Imbalance** We always use the same $\alpha$ for all weighted focal loss. Besides considering the label frequency, we also consider the performance gap between previous reported $F_1$. We

| Code | Count | Description | Examples |
|---|---|---|---|
| MIA | 3869 | Group of MI Adherent codes : Affirm(AF); Reframe(RF); Emphasize Control(EC); Support(SU); Filler(FI); Advise with permission(ADP); Structure(ST); Raise concern with permission(RCP) | "You've accomplished a difficult task." (AF)<br>"Its your decision whether you quit or not" (EC)<br>"That must have been difficult." (SU)<br>"Nice weather today!" (FI)<br>"Is it OK if I suggested something?" (ADP)<br>"Let's go to the next topic" (ST)<br>"Frankly, it worries me." (RCP) |
| MIN | 1019 | Group of MI Non-adherent codes: Confront(CO); Direct(DI); Advise without permission(ADW); Warn(WA); Raise concern without permission(RCW) | "You hurt the baby's health for cigarettes?" (CO)<br>"You need to xxx." (DI)<br>"You ask them not to drink at your house." (ADW)<br>"You will die if you don't stop smoking." (WA)<br>"You may use it again with your friends." (RCW) |

Table 1: Label distribution, description and exmaples for MIA and MIN

| Label | Categorizing | | | Forecasting | | |
|---|---|---|---|---|---|---|
| | P | R | $F_1$ | P | R | $F_1$ |
| FN | 92.5 | 86.8 | 89.6 | 90.8 | 80.3 | 85.2 |
| CT | 34.8 | 44.7 | 39.1 | 18.9 | 28.6 | 22.7 |
| ST | 28.2 | 39.9 | 33.1 | 19.5 | 33.7 | 24.7 |
| FA | 95.1 | 94.7 | 94.9 | 70.7 | 73.2 | 71.9 |
| RES | 50.3 | 61.3 | 55.2 | 20.1 | 18.8 | 19.5 |
| REC | 52.8 | 55.5 | 54.1 | 19.2 | 34.7 | 24.7 |
| GI | 74.6 | 75.1 | 74.8 | 52.8 | 67.5 | 59.2 |
| QUC | 80.6 | 70.4 | 75.1 | 36.2 | 24.3 | 29.1 |
| QUO | 85.3 | 81.2 | 83.2 | 27.0 | 11.8 | 16.4 |
| MIA | 61.8 | 52.4 | 56.7 | 27.0 | 10.6 | 15.2 |
| MIN | 27.7 | 28.5 | 28.1 | 17.2 | 10.2 | 12.8 |

Table 2: Performance of our proposed models with respect to precision, recall and $F_1$ on categorizing and forecasting tasks for client and therapist codes

choose to balance weights $\alpha$ as $\{1.0,1.0,0.25\}$ for CT,ST and FN respectively, and $\{0.5, 1.0, 1.0, 1.0, 0.75, 0.75,1.0,1.0\}$ for FA, RES, REC, GI, QUC, QUO, MIA, MIN. As shown in Table 5, we report our ablation studies on cross-entropy loss, weighted cross-entropy loss, and focal loss. Besides the fixed weights, focal loss offers flexible hyperparameters to weight examples in different tasks. Experiments shows that except for the model $\mathcal{C}^T$, focal loss outperforms cross-entropy loss and weighted cross entropy.

## References


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

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*.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.

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*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Michael Tanana, Kevin A Hallgren, Zac E Imel, David C Atkins, and Vivek Srikumar. 2016. A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of substance abuse treatment*, 65:43–50.

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 *Proceedings of the 2016 Conference of the International Speech Communication Association INTERSPEECH*, pages 908–912.

| Model | Embedding | macro | FN | CT | ST | macro | FA | RES | REC | GI | QUC | QUO | MIA | MIN |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{C}$ | ELMo | 53.9 | 89.6 | **39.1** | **33.1** | **65.4** | **95.0** | **55.7** | **54.9** | **74.2** | **74.8** | **82.6** | **56.6** | **29.7** |
| | $\text{ELMo}_{\text{psyc}}$ | 46.9 | 88.9 | 27.5 | 24.3 | 64.2 | 94.9 | 53.3 | 53.3 | 75.8 | 74.8 | 82.2 | 56.1 | 23.5 |
| | Glove | 50.6 | **89.9** | 33.4 | 28.6 | 62.2 | 94.6 | 53.7 | 54.2 | 70.3 | 70.0 | 79.1 | 54.7 | 20.9 |
| | $\text{Glove}^{\text{pysc}}$ | 47.4 | 88.4 | 23.9 | 30.0 | 63.4 | 94.9 | 54.7 | 52.8 | 75.2 | 71.4 | 80.8 | 53.6 | 23.5 |
| $\mathcal{F}$ | ELMo | **44.3** | **85.2** | **24.7** | 22.7 | **31.1** | 71.9 | 19.5 | **24.7** | **59.2** | 28.3 | **17.7** | 15.9 | 9.0 |
| | $\text{ELMo}_{\text{psyc}}$ | 43.8 | 84.0 | 22.4 | 25.0 | 29.1 | **73.5** | 15.5 | 24.3 | 59.1 | **29.1** | 9.5 | 12.1 | 10.1 |
| | Glove | 42.7 | 83.9 | 21.0 | 23.1 | 30.0 | 72.8 | **20.8** | 23.7 | 58.2 | 26.2 | 14.5 | 14.5 | 9.6 |
| | $\text{Glove}^{\text{pysc}}$ | 43.6 | 81.9 | 23.3 | **25.7** | 30.8 | 72.1 | 19.7 | 24.4 | 57.3 | 28.9 | 13.7 | **17.8** | **23.5** |

Table 3: Ablation study for our proposed model with embeddings trained on the psychotherapy corpus.

| Ablation | Options | CT | ST | R@3 | FA | RES | REC | GI | QUC | QUO | MIA | MIN |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| history size | 1 | 17.2 | 15.1 | 66.4 | 59.4 | 12.6 | 9.0 | 44.6 | 16.3 | 14.8 | 11.9 | 4.1 |
| | 4 | 16.8 | 22.6 | 75.3 | 71.4 | 15.6 | 21.1 | 57.1 | **29.3** | 11.0 | 11.2 | 14.4 |
| | $8^*$ | 24.7 | 22.7 | **77.0** | **72.8** | **20.8** | 23.1 | 58.1 | 28.3 | **17.7** | 15.9 | 9.0 |
| | 16 | 23.9 | 20.7 | 76.5 | 71.2 | 13.7 | 24.1 | **58.5** | 25.9 | 9.7 | 16.2 | 12.7 |
| word attention | GMGRU | 14.0 | **23.2** | 75.7 | 71.7 | 14.2 | 23.0 | 57.5 | 26.5 | 8.0 | 15.4 | 11.6 |
| | $\text{GMGRU}_{4h}$ | 19.1 | 22.9 | 76.3 | 71.3 | 12.1 | 23.3 | 58.1 | 24.5 | 12.6 | 11.7 | 14.0 |
| sentence attention | $-\ \text{SELF}_{42}$ | **24.9** | 22.5 | 76.0 | 71.4 | 12.7 | 24.9 | 58.3 | 28.8 | 5.9 | **17.4** | 9.7 |
| | $\setminus\ \text{ANCHOR}_{42}$ | 22.9 | 22.9 | 76.2 | 72.2 | 15.5 | **24.6** | 59.5 | 27.1 | 7.7 | 16.3 | 8.3 |
| | $+\ \text{GMGRU} \setminus \text{ANCHOR}_{42}$ | 6.8 | 23.4 | 76.9 | 70.8 | 8.0 | 24.5 | 58.3 | 24.6 | 10.6 | 14.9 | **12.1** |

Table 4: Ablation on forecasting task on both client and therapist code. $*$ row are results of our best forecasting model $\mathcal{F}_C$, and $\mathcal{F}_T$. $\setminus$ means substitute anchor attention with self attention. +GMGRU $\text{ANCHOR}_{42}$ means using word-level attention and achor-based sentence-level attention together.

| **Loss** | **Client** | | | **Therapist** | | | | |
|---|---|---|---|---|---|---|---|---|
| | $F_1$ | CT | ST | $F_1$ | RES | REC | MIA | MIN |
| $\mathcal{C}^{\text{ce}}$ | 47.0 | 28.4 | 22.0 | 60.9 | 54.3 | 53.8 | 53.7 | 4.8 |
| $\mathcal{C}^{\text{wce}}$ | 53.5 | 39.2 | 32.0 | 65.4 | 55.7 | 54.9 | 56.6 | 29.7 |
| $\mathcal{C}^{\text{fl}}$ | 53.9 | 39.1 | 33.1 | 65.4 | 55.7 | 54.9 | 56.6 | 29.7 |
| $\mathcal{F}^{\text{ce}}$ | 42.1 | 17.7 | 18.5 | 26.8 | 3.3 | 20.8 | 16.3 | 8.3 |
| $\mathcal{F}^{\text{wce}}$ | 43.1 | 20.6 | 23.3 | 30.7 | 17.9 | 25.0 | 17.7 | 10.9 |
| $\mathcal{F}^{\text{fl}}$ | 44.2 | 24.7 | 22.7 | 31.1 | 19.5 | 24.7 | 15.2 | 12.8 |

Table 5: Abalation study of different loss function on categorizing and forecasting task. Based on our proposed model for our four settings, we compared our best model with crossentropy loss(ce), $\alpha$ balanced cross-entropy(wce) and focal loss. Here we only report the macro $F_1$ for rare labels and the overall macro $F_1$. $\gamma = 1$ is the best for both the model $\mathcal{C}_C$ and $\mathcal{F}_C$, while $\gamma = 0$ is the best for $\mathcal{C}_T$ and $\gamma = 3$ for $\mathcal{F}_T$. Worth to mention, when $\gamma = 0$, the focal loss degraded into $\alpha$-balanced crossentropy, that first two rows are the same for therspit model.