INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

A Deep Learning Approach to Modeling Empathy in Addiction Counseling

James Gibson[1], *Doğan Can*[1], *Bo Xiao*[1], *Zac E. Imel*[2],
David C. Atkins[3], *Panayiotis Georgiou*[1], *Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA, USA
[2]Department Educational Psychology, University of Utah, Salt Lake City, UT, USA
[3]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA
[1]sail.usc.edu, [2]zac.imel@utah.edu, [3]datkins@u.washington.edu

Abstract

Motivational interviewing is a goal-oriented psychotherapy, employed in cases such as addiction, that aims to help clients explore and resolve their ambivalence about their problem. In motivational interviewing, it is desirable for the counselor to communicate empathy towards the client to promote better therapy outcomes. In this paper, we propose a deep neural network (DNN) system for predicting counselors' session level empathy ratings from transcripts of the interactions. First, we train a recurrent neural network mapping the text of each speaker turn to a set of task-specific behavioral acts that represent local dynamics of the client-counselor interaction. Subsequently, this network is used to initialize lower layers of a deep network predicting session level counselor empathy rating. We show that this method outperforms training the DNN end-to-end in a single stage and also outperforms a baseline neural network model that attempts to predict empathy ratings directly from text without modeling turn level behavioral dynamics.

Index Terms: behavioral signal processing, recurrent neural networks, word embedding, motivational interviews

1. Introduction

Modeling human communicative behaviors is a challenging undertaking. Machine learning offers possibilities for characterizing complex human behavior. Constructs of interest that are characterized, and learned, from human behavioral data are inherently multi-instance and multi-label, which presents new challenges and opportunities for researchers. Traditionally, such data are manually analyzed and studied by humans. This approach is costly and time consuming, which prompts the search for computational methods that can support and augment these efforts. Advances in machine learning present attractive avenues for behavioral analysis and modeling, both in enabling efficient means for computing desired behavioral constructs of interest, and in enabling discovery of new ones [1].

Motivational interviewing (MI) is a client-centered psychotherapy, which aims to help clients make behavioral changes through examination and resolution of ambivalence. Human communicative behaviors are especially important in MI, where the counselors' ability to express these behaviors can be vital for positive client outcomes. Researchers in the domain are interested in measures for relating counselor and client behaviors to counselor skill as well as intervention success [2, 3]. Furthermore there is considerable interest in how these measures directly relate to the spoken language of the counselor and client [4, 5].

Empathy is one behavior that has been of special interest in MI research as it is strongly associated with a counselor's ability to establish a rapport with their clients [6]. In the MI context *empathy*[1] is defined as, "the extent to which the therapist understands and/or makes an effort to grasp the clients perspective" [7]. Xiao et al. proposed modeling therapist *empathy* in motivational interviews using natural language processing [8]. They subsequently, analyzed therapist *empathy* using several approaches including prosody, speech rate, and vocal entrainment [9, 10].

In addition to global (session level) measures such as *empathy*, there have been multiple studies in employing machine learning approaches to model local (utterance level) participant behaviors in MI. Can et al. compared using a conditional random field (CRF) for modeling utterance level behaviors (behavioral acts) with dialogue acts and then related predicted counts of these acts to *empathy* [11]. Tanana et al. proposed using recursive neural networks paired with maximum entropy Markov models (MEMMs) to predict statements by clients about changing or maintaining their addictive behaviors [12]. Tanana et al. also compared recursive neural network models to discrete sentence features and reported improved accuracy of predicting several utterance level behaviors when using the recursive neural network model [13].

We draw inspiration from these works to propose a system that treats the local (turn level) behavioral acts as an encoding for the global (session level) *empathy* rating. We propose a deep learning system which uses the manually annotated local behavioral codes to train a recurrent neural network (RNN) which learns a mapping from the client/counselor dialogue to these local behavioral acts. Subsequently, this network is used to initialize the lower layers of a deep network for predicting the global counselor *empathy* rating.

2. Motivational Interviewing Data

We use a corpus of motivational interviews collected from six independent clinical studies. These studies all focused on addiction counseling relating to alcohol, marijuana, and other drug abuse. Three of these studies were aimed at reducing alcohol abuse by young people (ARC, ESPSB, ESB21), one focused on marijuana abuse (iCHAMP) and one on poly-drug abuse (HMCBI) [4]. The data from these five studies includes 148 sessions comprised of only real patients. Additionally, the Context Tailored Training (CTT) data includes 200 sessions of both real (76) and standardized (124) patients [14]. Standardized patients

[1] Note: we italicize *empathy* to distinguish this specific operational definition from colloquial definitions.

http://dx.doi.org/10.21437/Interspeech.2016-554

Table 1: MISC Categories

Code	Category	Count
	Counselor	
ADP	Advise with permission	105
ADW	Advise w/o permission	598
AF	Affirm	1649
CO	Confront	187
DI	Direct	134
EC	Emphasize Control	133
FA	Facilitate	16296
FI	Filler	157
GI	Giving Information	15748
QUC	Closed Question	5276
QUO	Open Question	4562
RCP	Raise Concern with permission	4
RCW	Raise Concern w/o permission	42
REC	Complex Reflection	4703
RES	Simple Reflection	6354
RF	Reframe	19
ST	Structure	1223
SU	Support	642
WA	Warn	65
	Client	
C+/C-	Commitment	111/21
FN	Follow/Neutral	47491
R+/R-	Reason	3278/2828
O+/O-	Other	1788/1638
TS+/TS-	Taking Steps	133/51

are actors portraying patients struggling with addiction for the purpose of counselor training.

All the data were manually transcribed and segmented at the turn level and received session level behavioral coding according to the motivational interviewing treatment integrity (MITI) manual [7]. Subsequently, in 337 of the transcripts the talk turns were segmented into utterances and the utterances were assigned local-level behavioral codes according to the motivational interviewing skill code (MISC) manual [15].

The MITI manual defines session level behavioral codes, e.g., *empathy*, and counts of utterance level behaviors, such as *reflections* and *questions*. The session level behavioral codes are rated on a 1-7 Likert scale. For this study we binarize the Likert scale rating as ‘high’ ($>= 5$), and ‘low’ (< 5). This cutoff is motivated by the MITI manual which defined a score of 5 as the minimum acceptable score for a counselor to be considered ‘proficient’.

The MISC manual defines 28 utterance level behavioral codes for the counselor (19) and the client (9). In Table 1, we show the full list of MISC codes and the number of times they occur in the data. Many of these codes are very difficult to predict due to their sparsity in the data. For this reason, Can et al. proposed grouping the codes into categories [11]. They proposed 8 categories including: FA, GI, QUC, QUO, REC, RES, COU, and CLI. The COU category groups all infrequent counselor codes and the CLI category groups all client codes. We refer to the full MISC code set as MISC28 and the reduced code set as MISC8.

The dataset is separated into training and testing sets. These sets follow approximately a 2:1 training (228 sessions) to testing (109 sessions) ratio. We took care to balance the *empathy* rating distributions between the sets, while maintaining speaker independence. Because some counselors appear in multiple sessions, all sessions from a particular counselor were assigned to the same train/test split. Furthermore, all standardized patient sessions were assigned to the train set. This was done because there are only three unique standardized patient stories, so the language is likely very similar between many of these sessions. We show the *empathy* distribution for the train and test sets in Figure 1.

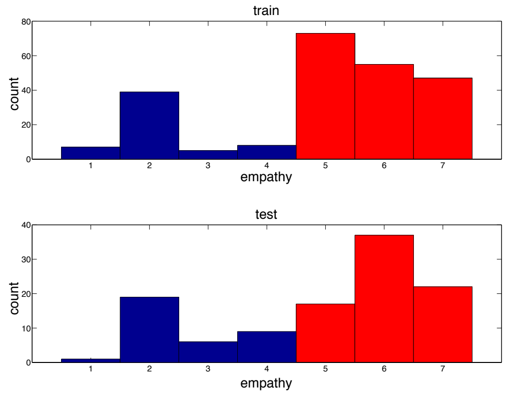

Figure 1: *Empathy* distribution of the train and test sets

3. Methodology

We learn word embeddings to represent the language use of the counselor and client using a continuous bag of words model [16]. Each word, w, is represented by an M-dimensional dense vector, V_w. Each turn in a session is represented by the average of the word vectors belonging to the words in that turn, i.e,

$$X_t = \frac{1}{|W_t|} \sum_{w \in W_t} V_w, \qquad (1)$$

where W_t is the set of words in turn t. Every session is now represented by a sequence of turn vectors, $X^i = \{X_t^i\}_{t=0}^{T^i}$, where T^i is the number of turns in session i. An additional indicator variable is appended to each vector to identify whether the turn belongs to the counselor or client.

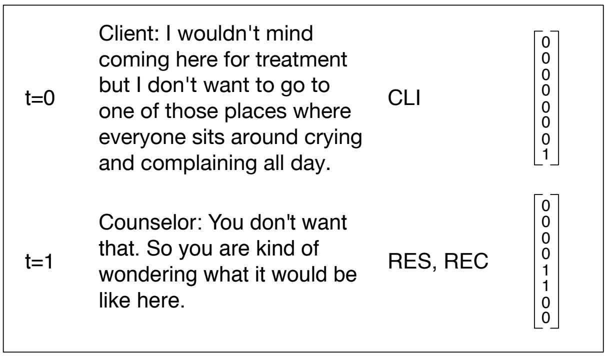
t=0
Client: I wouldn't mind coming here for treatment but I don't want to go to one of those places where everyone sits around crying and complaining all day.
CLI
0 0 0 0 0 0 0 1
t=1
Counselor: You don't want that. So you are kind of wondering what it would be like here.
RES, REC
0 0 0 0 1 1 0 0

Figure 2: Example MISC8 encoding.

Each turn has an associated L-dimensional k-hot target local behavior vector, Y_t^i, representing all k MISC codes that occur in that turn. We use either $L = 8$ or $L = 28$ depending on the cardinality of the MISC code set we are working with

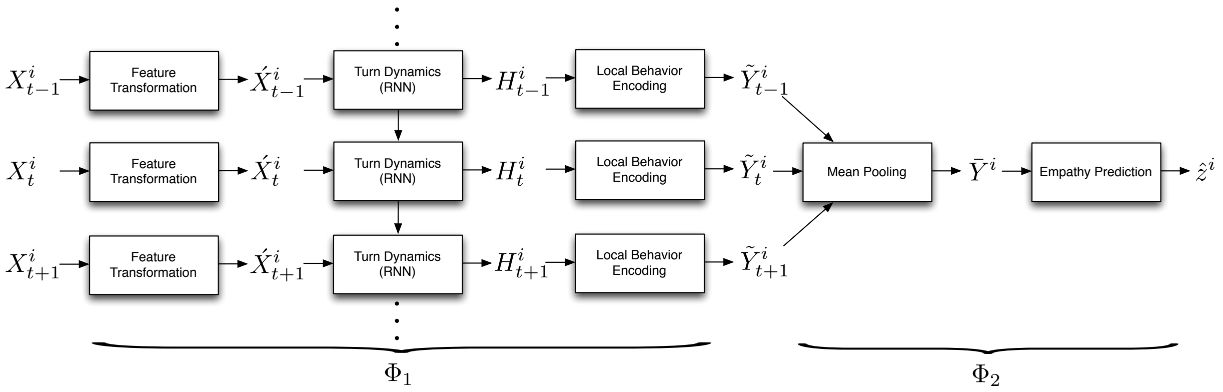

Figure 3: Proposed System Overview.

(see Section 2). An example of k-hot MISC8 vector is given in Figure 2.

We train the proposed system in two stages: the first maps from the turn vectors, X^i, to the k-hot representation of the local behaviors, Y^i; the second maps from the estimated local behaviors to the binarized session *empathy* score, z^i. We refer to the first stage as the encoder as it maps from the larger dimensional feature representation to a low dimensional representation of local behaviors. We refer to the second stage as the decoder as it attempts to predict the session *empathy* score from the estimated local behavioral encodings.

In the first stage, the turn vector sequences are input to a dense feedforward layer to allow for feature interactions. The transformed turn vectors, $\acute{X}_t^i$ (M+1-dimensional), are then input to an RNN which learns the dynamics of the turn vector sequence. The output of the RNN, H_t^i (M+1-dimensional), is input to a feedforward layer which maps to the local behavioral encoding, Y_t^i, an L-dimensional vector. We train this stage in a supervised manner using the reference local behavior vectors as the multi-label targets. We use a sigmoid activation function on each output of the encoding layer and the system is trained to minimize the average of the binary cross-entropy between the reference and predicted outputs as given by:

$$E_1 = -\sum_{i=1}^{N}\sum_{t=1}^{T^i}\sum_{l=1}^{L} Y_t^i(l)\cdot\log\left(\hat{Y}_t^i(l)\right) + \left(1-Y_t^i(l)\right)\cdot\log\left(1-\hat{Y}_t^i(l)\right), \qquad (2)$$

where,

$$\hat{Y}_t^i = \sigma(\Phi_1(X_t^i)), \qquad (3)$$

and Φ_1 is encoder. We use long short-term memory (LSTM) RNNs [17] to address the vanishing gradient problem that arises while training traditional RNNs [18].

In the second stage, we take the output of the first stage prior to the sigmoid activation, $\tilde{Y}_t^i = \Phi_1(X_t^i)$, as input. This encoding layer is averaged across turns and input to a feedforward layer that predicts the session's *empathy* rating z^i. We use the sigmoid activation function and binary cross-entropy, as given by:

$$E_2 = -\sum_{i=1}^{N} z^i\cdot\log(\hat{z}^i) + (1-z^i)\cdot\log(1-\hat{z}^i). \qquad (4)$$

where,

$$\hat{z}^i = \sigma(\Phi_2(\tilde{Y}^i)), \qquad (5)$$

Φ_2 is the decoder and $\tilde{Y}^i = \{\tilde{Y}_t^i\}_{t=1}^{T^i}$. We train the proposed system end-to-end, i.e. back-propagating the error from the *empathy* prediction layer back to the turn vector input layer. An overview of the proposed system is shown in Figure 3.

As a baseline model, we train a feedforward neural network with X^i as input to predict the *empathy* scores z^i. This model has a single fully connected layer ($X_t^i \rightarrow \acute{X}_t^i$) followed by averaging across all session turns ($\bar{X}^i = \frac{1}{T^i}\sum_{t=1}^{T^i}\acute{X}_t^i$), which is then input to the output layer, with sigmoid activation ($\bar{X}^i \rightarrow \hat{z}^i$).

We also predict *empathy* using the reference local behavioral representations, Y^i using the same network topology as the baseline, i.e., ($Y_t^i \rightarrow \acute{Y}_t^i \rightarrow \bar{Y}^i \rightarrow \hat{z}^i$). This serves to show how much information we can gain about the annotated global behavior from the annotated local behaviors.

4. Experiments and Results

We use the word2vec software to learn 300-dimensional word embedding vectors from the counselor and client transcripts [16, 19]. All neural network configurations are learned using Keras [20] with Theano [21] as the back-end.

We perform z-normalization on all features using the empirical mean and standard deviation of the training set. The turns from training sessions were segmented into sequences of 10 turns which overlapped by 50%. The data was shuffled and divided into 32 batches. Ten percent of the training data was randomly assigned to a validation set for training. During each training epoch the batches were shuffled to avoid overfitting to any particular batch. All layer weights were initialized with Glorot uniform initialization [22]. Ten percent dropout was applied to the output of each layer to avoid overfitting. The system was trained in 100 epochs (training was terminated early if the validation loss did not improve in three consecutive epochs) with the model from the epoch giving the minimum loss on the validation set being retained. The training procedure was optimized using the ADAM algorithm [23].

In Table 2, we show recall, precision, and the F1-score of the first stage, averaged over the result of each output target. The MISC8 representation is much more robustly predicted as it is a simplified version which only focuses on learning the most frequently occurring codes.

In Table 3, we show unweighted average recall (UAR) of predicting the global behavior, *empathy*. In this table, 'baseline' refers to the model that directly predicts *empathy* from the turn vectors and 'reference' refers to the model that predicts *empa-*

Table 2: MISC Prediction.

code	recall	precision	F1-score
MISC8	0.617	0.675	0.643
MISC28	0.228	0.348	0.258

thy from the reference MISC labels assigned to each turn. The proposed system was trained in two stages as described in the Section 3. This is referred to as 'pre-training' in the table. For reference we also include the results from training the full system end-to-end in a single stage, i.e., back-propagating the error from the *empathy* prediction layer back to the turn vector input layer without first learning the mapping from the turn vectors to local behaviors.

Table 3: *Empathy* Prediction.

model	L	UAR (%)
baseline	N/A	71.8
reference	8	73.6
	28	79.6
proposed system w/o pre-training	8	65.0
	28	62.9
proposed system w pre-training	8	78.6
	28	72.9

The reference model using all 28 MISC codes produced the highest UAR (79.6%) of all the prediction models. This demonstrates that the local behavioral codes carry important information about the global *empathy* code. The MISC8 reference model also gave better performance than the baseline model (73.6% vs. 71.8%), so while it does not make full use of the local codes the reduced set still carries important information for the *empathy* prediction task.

The proposed system, when trained in a single stage, fails to match the performance of the shallow baseline system (65.0/62.9% vs. 71.8%). This result suggests that there is not enough data to train a system of this depth without any supervision of the intermediary layers. When pre-training the encoder of the proposed system, we see improved performance over the baseline for both the 8 and 28 dimensional encoding layers (78.6% and 72.9%, respectively). Interestingly, we see a reversed order of performance between the 8 and 28 dimensional encoding layers compared to the predictions of the reference model. This is likely due to the difficulty of making accurate predictions of the 28 MISC codes including due to increased data sparsity issues. The relatively stronger performance of predicting the MISC8 labels gives a better initialization for the deep network. By back-propagating the error through the entire system with the first stage initialization, the deep system with 8 dimensional encoding layer achieves better performance than the reference system with reference MISC8 input (78.6% vs. 73.6%). This is most likely due to the deep system being allowed to learn from both the turn level language dynamics as well as local behavioral acts.

5. Conclusions and Future Work

In this paper, we presented a deep neural network system to predict counselor *empathy* from MI session transcripts. We demonstrated that by training the system in two stages, using local behavioral acts as supervision for the first stage, we are able to outperform a baseline shallow neural network predicting counselor *empathy* rating directly from input turn vectors.

In the future, we would like to model turns as sequences of word vectors rather than a simple average of word vectors. The current turn representation does not retain information about the relative order of words in a turn, only which words appeared. The order of words can be important for inferring meaning and thus improving this aspect of the model will likely lead to increased performance.

We also would like to improve the first stage of the training, i.e., the mapping from turns to local behavioral acts. While the outputs of the presented system share weights and thus are implicitly correlated, the loss function does not explicitly take into account the correlations between the labels (local codes). We plan to explore multi-labeling loss functions to address this issue [24].

A natural extension is to add an attention mechanism to the proposed encoder-decoder network [25]. An attention mechanism would allow for the system to assign different weights to different turns as some turns may carry more information about the *empathy* rating. By giving more attention to these behaviorally salient turns, we may achieve increased performance as well as increased interpretability.

Additionally, we are interested in using this system with text generated from MI session audio using automatic speech recognition (ASR), as well as direct audio derived features. In addition to the 348 MI sessions in the presented corpora, which were manually transcribed and annotated, we also have access to 1,384 sessions with global behavioral ratings but without the associated transcriptions or MISC annotations. A system that does automatic segmentation and ASR would be able to automatically provide behavioral codes for a therapy session without manual transcription. Our current "sound2code" system uses n-gram and maxent language models, lattice re-scoring, and support vector regression to make MITI predictions from ASR derived text [26]. We believe the proposed deep learning model could augment the existing system.

6. Acknowledgments

This work was supported by the NSF, NIH, and DOD.

7. References

[1] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, 2012.

[2] J. Gaume, G. Gmel, M. Faouzi, and J.-B. Daeppen, "Counselor skill influences outcomes of brief motivational interventions," *Journal of Substance Abuse Treatment*, vol. 37, no. 2, pp. 151–159, 2009.

[3] J. McCambridge, M. Day, B. A. Thomas, and J. Strang, "Fidelity to motivational interviewing and subsequent cannabis cessation among adolescents," *Addictive Behaviors*, vol. 36, no. 7, pp. 749–754, 2011.

[4] D. C. Atkins, M. Steyvers, Z. E. Imel, and P. Smyth, "Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification," *Implementation Science*, vol. 9, no. 1, p. 49, 2014.

[5] S. P. Lord, E. Sheng, Z. E. Imel, J. Baer, and D. C. Atkins, "More than reflections: Empathy in motivational interviewing includes language style synchrony between therapist and client," *Behavior Therapy*, 2014.

[6] R. Elliott, A. C. Bohart, J. C. Watson, and L. S. Greenberg, "Empathy." *Psychotherapy*, vol. 48, no. 1, p. 43, 2011.

[7] T. Moyers, T. Martin, J. Manuel, W. Miller, and D. Ernst, "The motivational interviewing treatment integrity (MITI) code: Version 2.0. university of new mexico, center on alcoholism," *Substance Abuse and Addictions (CASAA)*, vol. 2007, 2003.

[8] B. Xiao, P. G. Georgiou, and S. Narayanan, "Analyzing the language of therapist empathy in motivational interview based psychotherapy," in *Asia-Pacific Signal and Information Processing Association*, 2012, pp. 1–4.

[9] B. Xiao, P. G. Georgiou, Z. E. Imel, D. C. Atkins, and S. Narayanan, "Modeling therapist empathy and vocal entrainment in drug addiction counseling." in *INTERSPEECH*, 2013, pp. 2861–2865.

[10] B. Xiao, Z. E. Imel, D. C. Atkins, P. G. Georgiou, and S. S. Narayanan, "Analyzing speech rate entrainment and its relation to therapist empathy in drug addiction counseling," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[11] D. Can, D. C. Atkins, and S. S. Narayanan, "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*, 2015.

[12] M. Tanana, K. Hallgren, Z. Imel, D. Atkins, P. Smyth, and V. Srikumar, "Recursive neural networks for coding therapist and patient behavior in motivational interviewing," *NAACL HLT 2015*, p. 71, 2015.

[13] M. Tanana, K. A. Hallgren, Z. E. Imel, D. C. Atkins, and V. Srikumar, "A comparison of natural language processing methods for automated coding of motivational interviewing," *Journal of substance abuse treatment*, 2016.

[14] J. S. Baer, E. A. Wells, D. B. Rosengren, B. Hartzler, B. Beadnell, and C. Dunn, "Agency context and tailored training in technology transfer: A pilot evaluation of motivational interviewing training for community counselors," *Journal of Substance Abuse Treatment*, vol. 37, no. 2, pp. 191–202, 2009.

[15] W. R. Miller, T. B. Moyers, D. Ernst, and P. Amrhein, "Manual for the motivational interviewing skill code (MISC)," *Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*, 2003.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *Neural Networks, IEEE Transactions on*, vol. 5, no. 2, pp. 157–166, 1994.

[19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[20] F. Chollet, "Keras," https://github.com/fchollet/keras, 2015.

[21] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio, "Theano: new features and speed improvements," Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

[22] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[23] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] M.-L. Zhang and Z.-H. Zhou, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1338–1351, 2006.

[25] K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1875–1886, 2015.

[26] B. Xiao, Z. E. Imel, P. G. Georgiou, D. C. Atkins, and S. S. Narayanan, "'rate my therapist': Automated detection of empathy in drug and alcohol counseling via speech and language processing," *PloS one*, vol. 10, no. 12, p. e0143055, 2015.