Multimodal Learning for Identifying Opportunities for Empathetic Responses

Leili Tavabi
University of Southern California
Institute for Creative Technologies
Los Angeles, CA, USA
ltavabi@ict.usc.edu

Kalin Stefanov
University of Southern California
Institute for Creative Technologies
Los Angeles, CA, USA
kstefanov@ict.usc.edu

Setareh Nasihati Gilani
University of Southern California
Institute for Creative Technologies
Los Angeles, CA, USA
sngilani@ict.usc.edu

David Traum
University of Southern California
Institute for Creative Technologies
Los Angeles, CA, USA
traum@ict.usc.edu

Mohammad Soleymani
University of Southern California
Institute for Creative Technologies
Los Angeles, CA, USA
soleymani@ict.usc.edu

ABSTRACT

Embodied interactive agents possessing emotional intelligence and empathy can create natural and engaging social interactions. Providing appropriate responses by interactive virtual agents requires the ability to perceive users' emotional states. In this paper, we study and analyze behavioral cues that indicate an opportunity to provide an empathetic response. Emotional tone in language in addition to facial expressions are strong indicators of dramatic sentiment in conversation that warrant an empathetic response. To automatically recognize such instances, we develop a multimodal deep neural network for identifying opportunities when the agent should express positive or negative empathetic responses. We train and evaluate our model using audio, video and language from human-agent interactions in a wizard-of-Oz setting, using the wizard's empathetic responses and annotations collected on Amazon Mechanical Turk as ground-truth labels. Our model outperforms a text-based baseline achieving F1-score of 0.71 on a three-class classification. We further investigate the results and evaluate the capability of such a model to be deployed for real-world human-agent interactions.

ICMI '19, October 14–18, 2019, Suzhou, China

ACM ISBN 978-1-4503-6860-5/19/10…$15.00
https://doi.org/10.1145/3340555.3353750

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence; Information extraction; Neural networks; • **Human-centered computing** → Laboratory experiments.

KEYWORDS

multimodal sentiment, virtual human, empathy, machine learning, human behavior

ACM Reference Format:
Leili Tavabi, Kalin Stefanov, Setareh Nasihati Gilani, David Traum, and Mohammad Soleymani. 2019. Multimodal Learning for Identifying Opportunities for Empathetic Responses. In *2019 International Conference on Multimodal Interaction (ICMI '19), October 14–18, 2019, Suzhou, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340555.3353750

1 INTRODUCTION

Emotionally intelligent and embodied interactive agents are showing great promise for effectively augmenting human resources in different domains including health-care and education. To create a realistic and engaging experience, it is necessary for the agents to be receptive and responsive to the users' emotional needs. There has been a large body of work in multimodal recognition of sentiment and human emotions from online videos or interactive experiences [7, 8, 29]. Existing work have made notable progress towards sentiment recognition from vast online datasets. Nonetheless, despite the increasing attention towards emotionally intelligent and empathetic interactive companions, recognition of empathy has not been extensively explored due to limited amount of data and the complexity of defining ground-truth labels.

Empathy is defined as the ability to recognize, understand and react to emotions, attitudes and beliefs of others [1]. Automatic recognition of empathy, although similar to

sentiment, requires a different and more complex modeling. Recognition of opportunities for empathetic responses should include subjectivity while also accounting for the intensity of the sentiment to elicit empathetic responses. The threshold for expressing empathetic responses can vary from person to person and is also affected by inter-personal relationships and the context of the conversation. *"I am concerned about global warming."* and *"I lost my mother to cancer."* are expected to elicit different responses in terms of empathy.

Multimodal sentiment analysis relies on the perception of proxies of sentiment or affect from different views including verbal content or spoken words, emotional tone of speech and facial expressions. In this work, building upon the work on multimodal sentiment analysis, we propose a multimodal machine learning framework for identifying opportunities for empathetic responses during human-agent conversations. To this end, we analyzed interactions between an agent and a user during a semi-structured interview probing symptoms of mental health disorders such as depression. During the interview, the agent asks a set of questions, where each question is possibly followed by shorter follow-up questions with respect to the user's previous responses. Our developed model determines when the agent needs to express empathy and with what polarity, *i.e.*, a positive or negative empathetic response. We focus on the prediction of empathy in an uncontrolled environment with real-world users, throughout the human-agent dialogue interaction.

The problem is therefore formulated as a three-class classification of positive, negative or no response using verbal, acoustic and visual modalities. Each modality is mapped to a representation which is used to recognize the classes. We evaluated the unimodal and multimodal recognition results with two sets of labels, one consisting of the real-time judgments of the experimenters and one according to the judgments of the independent observers. Identifying such moments will enable the agent to provide an empathetic response such as "I'm sorry" or "that's great" when necessary.

The major contributions of this work include:

- An analysis of verbal and nonverbal behaviors prompting empathetic responses.
- Providing a machine learning framework for identifying empathetic opportunities in an uncontrolled dyadic interaction with real-world users.
- An analysis of different strategies for creating ground-truth labels for empathetic responses.

2 RELATED WORK

The development of emotionally intelligent and empathetic agents have been a long-standing goal of AI. Bickmore [5] showed how embodied agents can employ empathy to form better social relationships. Brave *et al.* [6] shows that empathetic emotions lead to greater likeability and trustworthiness of the agent. Existing work have mostly examined empathetic interactions through game-playing contexts [4, 6, 23].

Others have looked at prediction of counselors' empathy measures in domains like motivational interviewing [34, 35]. They have used ratings of empathy as means of evaluating psychotherapy sessions and counselor performance. Clavel and Callejas [10] surveyed sentiment analysis and its applications to human-agent interaction. They found that the existing sentiment analysis methods deployed in human-agent interactions are not designed for socio-affective interactions. Hence, they recommend building systems that can support socio-affective interactions in addition to enhancing engagement and agent likability.

Sentiment analysis usually focuses on recognizing the polarity of sentiment expressed towards an entity [32]. Learning empathetic opportunities in interactive systems requires more than mere recognition of polarity, since empathetic responses are in response to personal misfortunes or successes and not just any emotionally charged utterance.

Recent multimodal sentiment analysis approaches use deep neural networks trained and evaluated on social media videos to detect sentiment. Zadeh et al. [37] used a Tensor Fusion Network to model intra-modality and inter-modality dynamics in multimodal sentiment analysis. Their tensor fusion network consists of modality embedding sub-networks, a tensor fusion layer modeling the unimodal, bimodal and trimodal interactions using a three-fold Cartesian product from modality embeddings along with a final sentiment inference sub-network conditioned on the tensor fusion layer.

Hazarika *et al.* [20] propose a conversational memory network for emotion recognition in dyadic interactions, considering emotion dynamics. They use Gated Recurrent Units (GRUs) to model past utterances of each speaker into memories to leverage contextual information from the conversation history. Majumder et al. [24] models emotions in conversations by distinguishing individual parties throughout the conversation flow. They consider three major aspects in dialogue by modeling individual party states, context from the preceding utterances as well as the emotion of the preceding utterance by employing three GRUs. Their network feeds incoming utterances into two GRUs called Global GRU and party GRU to update the context and party states respectively. The global GRU encodes corresponding party information while encoding an utterance. By attending over the global GRU, the model represents information from all previous utterances and the speaker state. Depending on the context, information is updated and fed into the emotion GRU for emotion representation.

Existing work mostly leverage online datasets that benefit from large amounts of data [9, 27, 28, 30], or use highly

Figure 1: A participant and the virtual agent, Ellie.

curated offline datasets that adopt professional actors for predefined and highly expressive scenarios [25, 30, 33]. In this paper, we focus on real-world data obtained from people talking with a virtual agent in a semi-structured interview imitating a therapy session. This is an inherently challenging domain due to limited amount of real-world data with relatively lower expressiveness and unstructured spoken dialogue.

3 DATA

We use a portion of the Distress Analysis Interview Corpus - Wizard-of-Oz (DAIC-WOZ) for training and evaluating our method. DAIC-WOZ is a subset of DAIC that contains semi-structured interviews designed to support the assessment of psychological distress conditions such as depression and post-traumatic stress disorder (PTSD) [19]. The interviews were collected as part of an effort to create a virtual agent that conducts semi-structured interviews to identify verbal and nonverbal indicators of mental illness. The subset of the corpus examined in this work include the Wizard-Of-Oz interviews conducted by a virtual agent controlled by two trained human wizards in a separate room. In this two-wizard arrangement, one wizard controlled the the agent's verbal behavior while the other handled her nonverbal behavior. The interview was structured to start with a set of general rapport-building questions and continue to query potential symptoms of mental health such as quality of sleep. In this setup, a fixed set of top-level questions were provided to the wizard to be asked during the interview. In addition to asking the top-level questions, the wizard was provided with a finite repertoire of response options to act as a good listener by providing back-channels, empathy and continuation prompts [13] (see Figure 1).

Verbal and nonverbal behavior of participants were captured by a front-facing camera and head-worn microphone. In this work, we extract segments eliciting empathetic responses from the experiments by looking at the agent's expressions of empathy such as "I'm sorry to hear that." or "That sounds like a great situation.". In the segmented data, each instance consists of the participants' verbal and nonverbal (audiovisual) responses to each main question and the follow-up questions. Follow-up questions such as "Can you tell me more about that?" were asked to elicit further disclosure and encourage more elaborate responses. Example dialogue excerpts are shown in Table 1.

Table 1: Human-Agent dialogue excerpts with different empathy responses.

	Dialogue Excerpt
Negative	**A:** How have you been feeling lately? **H**: Um kind of uh I guess sorta sorta depressed generally **A:** Tell me more about that **H:** Uh just uh feeling tired and sluggish and um less less motivated and less interested in things **A:** I'm sorry to hear that.
Positive	**A:** What are you most proud of in your life? **H:** Uh I'm proud that I've come a long way from when I first moved out here I'm uh a lot more disciplined um I read a lot uh I do crosswords and I think I've I think I know what's important in life now and I'm more focused and going after what I want **A:** That's so good to hear.
None	**A:** What are somethings you wish you could change about yourself? **H:** Um I wish I could be taller I wish I could be more inclined to play basketball so I then become go to the NBA and be a millionaire I know that's all unrealistic but just answering honestly.

Due to the nature of the predefined semi-structured interview, the dialogue turns take minimal influence from the dialogue history and are therefore considered independently. The data is segmented into small time-windows consisting of the users' transcribed text, video and audio that have resulted in either positive, negative or no empathetic responses from the virtual agent.

Overall, we had 2185 data points extracted from conversations of 186 participants. The average length of the dialogue excerpts was 30.6 seconds, while the average number of turns per data point was 3.2 turns.

4 METHOD

Multimodal Feature Extraction

Textual Features. For text input, we use a pre-trained language representation model called Bidirectional Encoder Representations from Transformers (BERT) [14]. BERT has

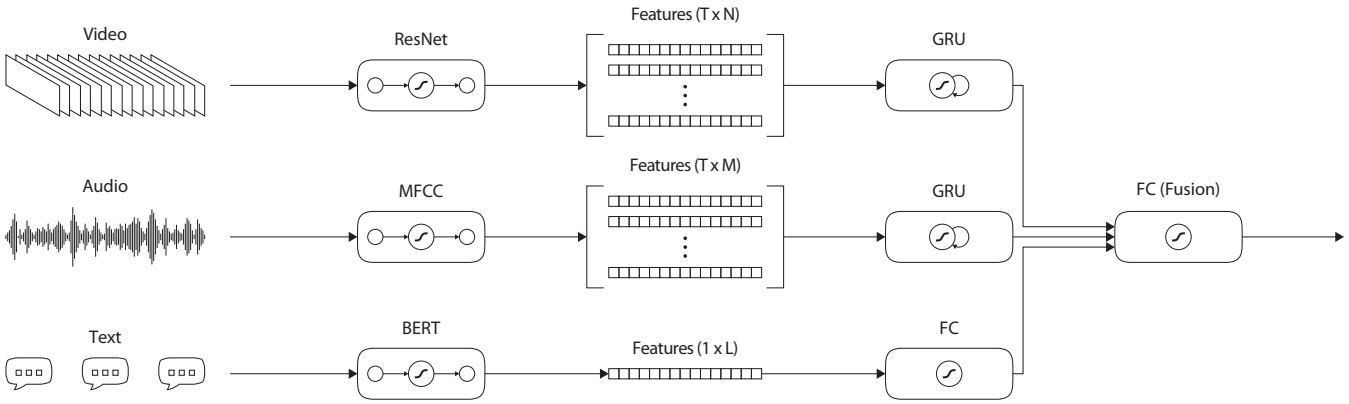

Figure 2: Multimodal static fusion.

substantially advanced the state-of-the-art in a number of natural language processing (NLP) tasks including sentiment analysis and question answering, which also makes it suitable for this task. We therefore used BERT as our text embedding model using only the participants' utterances from the dialogue excerpts. We avoid using the agent's utterances in the classification because of the unfair advantage it may provide to the recognition model. We took Uncased BERT-Base to obtain a single 768-dimension vector representation of the transcribed text per data entry [36]. BERT encodes the whole text sequence into a fixed-size vector, and unlike audiovisual modalities, the temporal dimension is latent in the text representation.

Audio Features. Two types of feature-sets were extracted for the representation of speech prosody: (i) the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) and (ii) Mel-frequency cepstral coefficients (MFCC), extracted using OpenSMILE [18].

eGeMAPS provides a set of acoustic features hand-selected by experts for their potential to detect affect in speech, and has been widely used in literature due to their performance, as well as theoretical significance [16]. This feature-set consists of 23 features such as fundamental frequency and loudness. MFCCs represent 13 band mel-frequency cepstral coefficients (MFCC) computed from audio signals from 25ms audio frames. MFCCs and their first and second order derivatives were extracted [17, 18] to obtain a temporal matrix of $T \times 39$ representation per data entry.

Visual Features. For the visual representation, we experimented with two different feature-sets: (i) 17 action units and 6 head pose features were extracted per frame using OpenFace [3] and (ii) face embedding obtained from a ResNet pre-trained model [21]. OpenFace is used to extract the intensity of facial action units, representing 17 action units based on the Facial Action Coding System (FACS) [15] along with head pose variations per frame, therefore providing a $T \times 23$ representation. For the face embedding, we extracted masked and aligned faces per frame using OpenFace [2] and fed it to ResNet-50, a convolutional neural network pre-trained on ImageNet [12], and extracted the representation from the penultimate layer, to obtain a $T \times 2048$ representation.

Ground-Truth Labels

Wizard judgments. We extracted the ground-truth labels from the empathetic and non-empathetic responses of the human-controlled virtual agent. The agent's responses are divided into three classes: negative empathy, positive empathy or no empathy. Negative empathetic responses include utterances such as "That sounds really hard" and "I'm sorry to hear that", positive empathy includes utterances like "That's so good to hear", "That sounds like a great situation", and no empathetic responses shows that the agent moved on to the next question or expressed fillers or back-channels without sentiment. By using these key phrases, we extracted the ground-truth labels for the three classes.

Mechanical Turk Ratings. To validate the wizard's empathetic responses, we collected labels via Amazon Mechanical Turk (MTurk). We recruited five raters per instance (257 unique participants), all from the United States to avoid language barriers. For each data point, the users were given the textual data, *i.e.*, the dialogue sequence and they were asked to select the proper categorical response toward the user at the end of each conversation. For further clarification, we provided example responses belonging to each category. Each assignment consisted of 20 tasks (data points) plus two control questions (with obvious responses) to eliminate raters that did not pay attention to the task and provided random answers. One control question contained an obviously devastating story about the participant's mother passing away while the other control question involved a very happy and inspiring story about the participant. We repeated the experiment on data points with wrong answers to either of

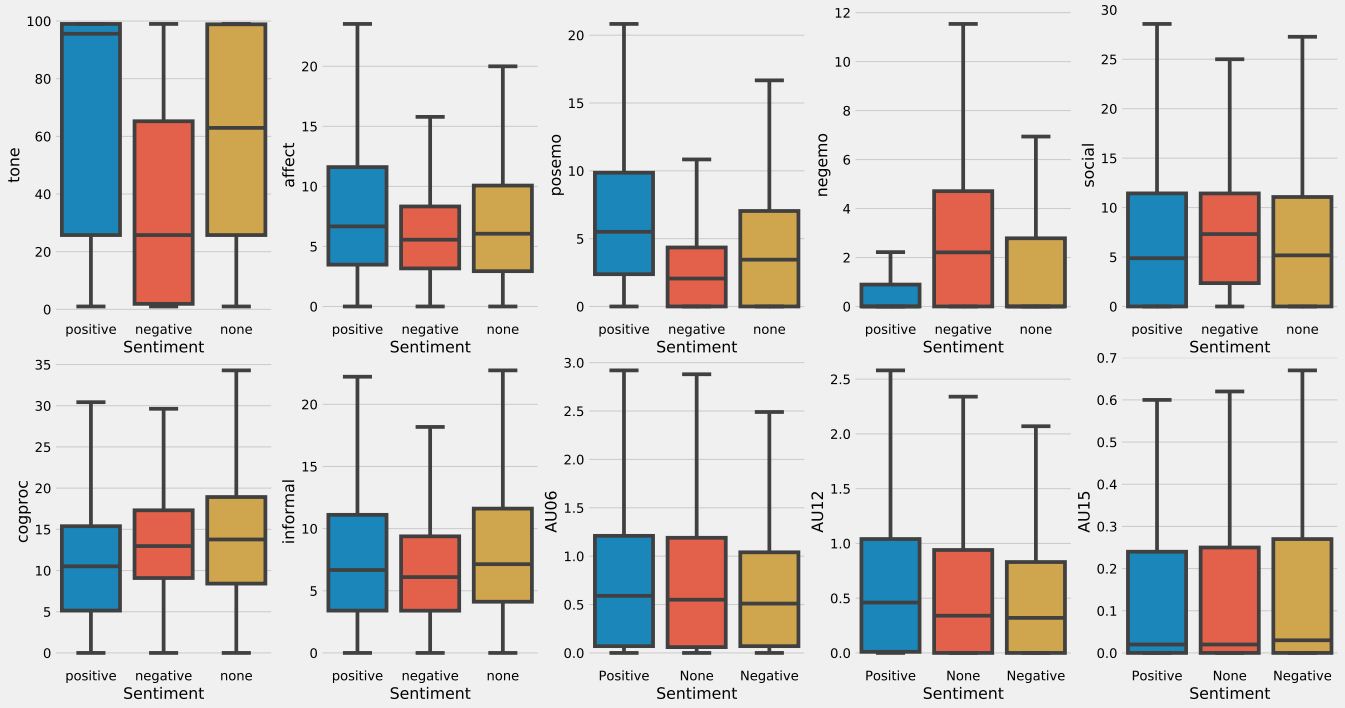

Figure 3: Box plots of verbal and nonverbal behavior with significant differences among different classes.

the control questions to obtain valid ratings. We additionally eliminated the instances where there was no majority vote among raters (7% of the data).

The Fleiss' kappa was calculated to measure inter annotator agreement for the entire data across five raters which showed fair agreement with $\kappa = 0.33$. A comparison between the majority vote of the MTurk raters and the wizard's responses, shows 58% agreement. More analysis indicates that the difference is mainly caused by MTurk raters annotating certain entries as either positive or negative where there was in fact no empathetic response by the wizard. This is likely the result of the raters looking at data entries independently and not as part of an entire dialogue. Therefore the wizard may not have expressed empathy where it was fit, to avoid redundancy of such expressions throughout the interaction. The low inter-rater agreement from MTurk annotations demonstrates the intrinsic complexity of the task, which speaks well to the nature of empathy as a social construct and the empathy level of the person expressing it. Furthermore, the task becomes more difficult due to the individual differences across the annotators with respect to their own personal experiences and self-identification with the user.

Table 2 shows the distribution of data across different classes. Throughout the experimentation, we evaluate and report the results for both sets of labels to address this difference between the sets of labels.

Table 2: Distribution of classes for two sets of labels

	Negative	Positive	None
Wizard	20.6%	40.6%	38.8%
MTurk	24.9%	46.0%	29.1%

Behavior Analysis

To study the verbal and nonverbal indicators associated with instances of behavior that elicit empathetic responses, we used interpretable features from each modality for investigating such associations. For vision, we used facial action units, for speech, we opted for eGeMAPS features and for language we used LIWC to gain a better understanding of the social predictive signals of empathy. Linguistic Inquiry and Word Count (LIWC) is a dictionary-based tool that generates scores along different dimensions including linguistic variables such as number of conjunctions and pronouns and affective and cognitive constructs [26].

After selecting a set of features, we ran one-way analysis of variance (ANOVA) and visually inspected the box plots of significant results ($p < 1E-5$). The behavioral features that stood out are shown in Figure 3. We could not observe any visible differences among audio features. The sentiment of language, tone, positive (posemo) and negative emotions

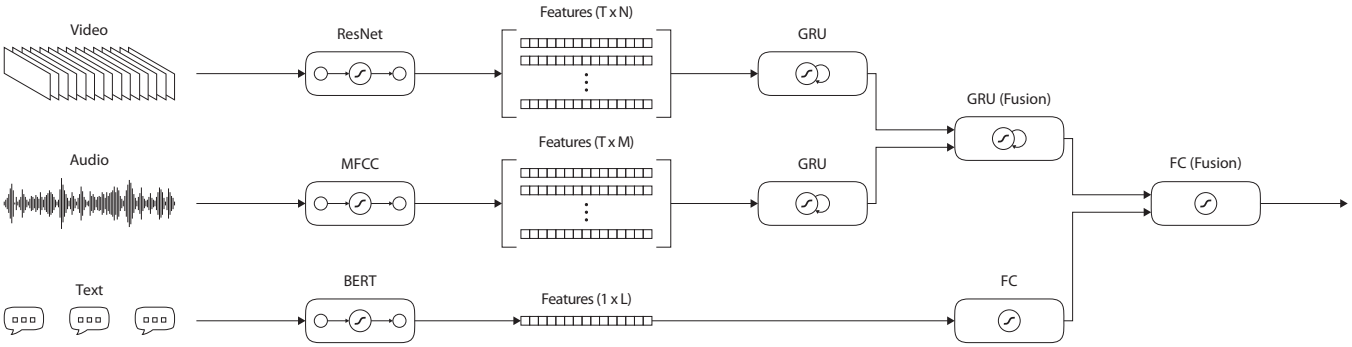

Figure 4: Multimodal RNN fusion.

(negemo), according to LIWC are strong indicators of sentiment for recognizing empathetic response opportunities. The language used in describing less pleasant situations is more formal which might show that participants were less comfortable sharing them. Social processes including mentioning family members was higher during the description of negative experiences, pointing toward interpersonal issues. Cognitive processes (cogproc) which involve describing causation, certainty and insight were lower for positive instances which demonstrate that the expressions of positive experiences were in simpler language.

Action units associated with positive expressions, AU06 (cheek raiser) and AU12 (lip puller), are strong indicators of positive sentiment. AU15 or lip corner depressor that is associated with sadness also showed stronger activation during negative instances. This demonstrates that visual features in addition to verbal behavior might be able to assist the recognition of sentiment for providing empathetic responses.

Model Architecture

Unimodal models. For every modality an encoder maps its input representations to a fixed-size vector or embedding. In unimodal classification, each of these encoders is then followed by a softmax layer for three class classification.

Language information is encoded with instance-based encoders. These encoders consist of a single fully connected (FC) layer of a fixed size. Sequences of audio and visual features were fed to a single layer gated recurrent unit (GRU) that maps the vision and speech representations to a fixed-size embedding, keeping only the last state. The obtained representations from unimodal encoders are followed by a softmax layer for classification. Additionally, we developed a multimodal model that fused the aforementioned encoders, described below.

Static fusion. In this architecture, features from different modalities are initially passed through unimodal encoders, and their resulting embeddings were concatenated and fed into a fully-connected fusion layer followed by a softmax classifier. The structure of this static fusion network is illustrated in Figure 2.

RNN Fusion. Similar to the static fusion model, the RNN fusion architecture initially produces unimodal embeddings for each modality. However, in case of vision and audio, with RNN encoders, the temporal embeddings learned through single-layer GRUs, are concatenated and fed into an RNN fusion layer consisting of a single-layer GRU. The text embedding is then concatenated with the output from the last state of the RNN fusion layer and fed to a single fully-connected layer for final fusion (static). The output is finally passed through a softmax classifier for three class classification. The RNN fusion network structure is shown in Figure 4.

Experimental Setup

In this work, we evaluate our methods on a dataset of 2185 instances of conversation excerpts from 186 participants. Given the size of the dataset at-hand, we opted for a simpler neural network architecture that can capture the patterns associated with empathetic responses while generalizing well. The model takes temporal audio and video input features per data entry and a single representation vector for text. We discard all data shorter than 1.5 seconds and apply random cropping of a 90-second window for long video and audio inputs (average length of the data is 90 seconds), during training. During evaluation a middle segment with max duration of 90 seconds is extracted.

For each modality, we designed an encoder network mapping the input feature space to a 128-d embedding space. In both architectures, video and audio inputs are fed separately into two 1-layer GRUs to obtain individual embeddings for both modalities. Only for ResNet due to the higher dimensionality of the original space, we added a 128-d fully connected layer after GRU. For textual data, the BERT vector representation is fed into a fully-connected layer to obtain a compact representation, reducing the feature dimensions from 768 to 128. The embeddings from all modalities are consistent across the two fusion networks. The two models

employ different fusion architectures: (i) static fusion model uses the concatenation of the three embeddings and feeds the multimodal representation vector to a fully-connected layer, with a dropout value of 0.2, to obtain a final vector of size three, containing the probabilities among three classes. A softmax classifier is then adopted to perform the classification (ii) RNN fusion model initially fuses the temporal video and audio sequences using a GRU of size 128 and then concatenates the bimodal representation with the text embedding. Similar to the static fusion network, the multimodal representation is fed to the fully-connected layer, with a dropout value of 0.2, obtaining the final probability vector on which a softmax classifier performs the classification. A cross-entropy loss is used in this setup with a weight vector, learned from the train set, to account for the data imbalance and the evaluation results are computed using micro F1-score. A 10-fold cross-validation has been used for training and evaluation of the dataset. We optimize the network using Adam, with a batch size of 32 and a learning rate of 10^{-4}. 20% of training data is held out in each iteration for validation, and the best performing model on the validation set is selected. In the case of multimodal models, the encoders and fusions layers are all trained jointly for 100 epochs.

Since there is no prior work whose results are directly comparable with our work, we compare our results against a text-based sentiment analysis method, given the similarities between our problem and classical sentiment analysis. For our text baseline, we use Valence Aware Dictionary and Sentiment Reasoner (VADER) which is a lexicon and rule-based sentiment analysis tool [22].

5 RESULTS AND DISCUSSION

To inform our design decisions for the multimodal networks, we initially trained and evaluated unimodal classifiers using different feature-sets. The results from unimodal classification, evaluated by micro F1-scores are shown in Table 3.

Unimodal classification results demonstrate the superiority of text in content representation and predictive power, exceeding performance from visual and audio modalities. This result is consistent with prior work on multimodal sentiment analysis [28, 37], and extenuated by the real-world setting and low expressiveness of this interactive scenario.

The multimodal networks are trained on the best performing feature-sets from each modality, meaning ResNet for video representation, MFCCs for audio and BERT for language. The audio representations had low predictive power for both MFCC and eGeMAPs on unimodal classifications, which may be the result of audio quality and recording. When training the models with MTurk annotations, the results from the multimodal networks show an increase in performance using the RNN fusion model, which speaks to the existing

Table 3: F1-scores for three-class classification.

	Features/Models	MTurk	Wizard
Audio	MFCC	**0.38**	0.36
	eGeMAPS	0.37	0.35
Video	AU+Pose	0.38	0.35
	ResNet	**0.46**	0.43
Text	BERT	**0.64**	0.61
Multimodal	Static Fusion	0.69	0.61
	RNN Fusion	**0.71**	0.61
Baseline	VADER - text	0.58	0.44

temporal inter-dynamics of audio and video captured by this network. The multimodal networks gain an overall advantage over the textual unimodal network which is the highest performing unimodal classifier in this task (see Table 3).

Our unimodal text classifier outperforms the text sentiment baseline. Using the recommended threshold on compound sentiment score, *i.e.*, 0.05 for VADER, a text-based sentiment analysis achieves $F1 = 0.58$ for MTurk labels and $F1 = 0.44$ for wizard labels. We also tested the sensitivity of the threshold value and found that the best possible results are only slightly different (see Figure 5). Hence, our text-based method using BERT comfortably outperforms VADER results which further validates our approach.

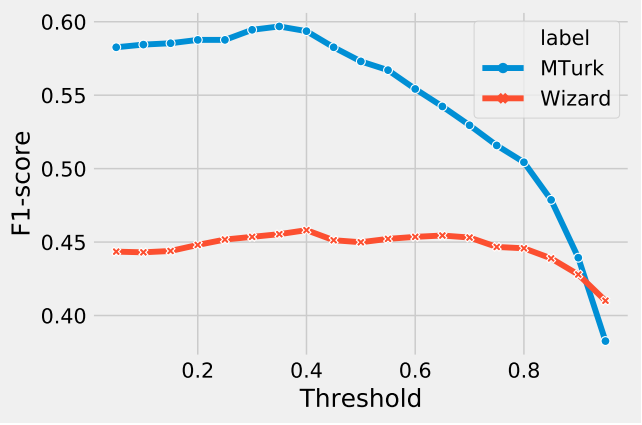

Figure 5: F1-scores of VADER sentiment analysis with different thresholds.

The results demonstrate that model predictions are higher when trained on MTurk labels for both multimodal and unimodal classifications. The aggregate of labels from five annotators provide higher reliability and potentially lower

between-person variability. Additionally, the wizard has an understanding of conversation context and may experience different inter-personal connections to the story or person that would affect the empathetic responses beyond the ability of our model.

The column-wise-normalized confusion matrices for RNN fusion model across wizard and MTurk ratings are shown in Table 4. The results show similar patterns for both labels and indicate that false predictions are mainly mis-classifications of either positive and negative responses with no empathy, *i.e.*, , prediction of positive/negative responses where none was necessary or predicting no empathy where positive empathy would have been a better response. To deploy such a system in real interactions, high precision in detection is necessary, as confusion of positive and negative responses will disrupt the interaction. Examples of the model's predictions on MTurk labels are shown in the Table 5. Instances like the second entry are dependent on the personalities and inter-personal relationships of the interlocutors. However, instances like the third entry can be disruptive to the interaction and require further attention.

Table 4: Confusion matrices (RNN fusion).

		Predictions		
		Negative	Positive	None
MTurk Labels	Negative	72.12%	4.15%	11.43%
	Positive	9.85%	78.44%	30.60%
	None	18.03%	17.41%	57.97%
Wizard Labels	Negative	49.65%	1.59%	12.02%
	Positive	14.24%	74.87%	30.80%
	None	36.11%	23.54%	57.18%

Table 5: Instances of RNN Fusion model's correct/incorrect predictions on MTurk labels (*Positive*, *Negative*, *None*).

Dialogue Excerpt	Prediction/Label
A: What got you to seek help? **H**: My mood was just not right I was always feeling down and depressed and lack of energy always wanting to sleep um lack of interest	Neg/Neg
A: What's your dream job? **H:** Designing for the movie industry **A:** How hard is that? **H:** Extremely so I never really pursued it	Non/Neg
A: What do you do when you're annoyed? **H:** When I'm annoyed you know I really don't get annoyed that much I just let it go it's not worth the pain and problems they could cause if I can't straighten out a problem let it go	Neg/Pos

6 CONCLUSIONS

In this paper, we reported on our efforts in automatic recognition of opportunities for providing empathetic responses. To this end, we labeled and analyzed a dataset of human-agent interactions in the context of a semi-structured interview. Our analysis demonstrated that facial expressions of emotions and verbal content are the important channels for recognizing such opportunities.

We developed and evaluated a deep neural network capable of multimodal learning of such opportunities. The best unimodal result was achieved by encoding language with a Transformer network (BERT) pre-trained on a large amount of data and performing classification. Fusing the verbal channel with facial expressions, our recurrent neural network fusion provided the best result of $F1 = 0.71$ which is comparable to the recent work on multimodal sentiment analysis [37].

Analysis on two sets of ground-truth labels from the experiments and independent observers, showed that empathy, similar to other social constructs, may suffer from indistinct boundaries that can be affected by inter-personal relationships and individuals' personalities.

As part of future work, to prepare this framework for real-time use, we will optimize the current model for precision on positive and negative classes, while assigning higher emphasis on instances with unanimous labels among annotators. Ultimately, such models should be able to model individual differences by choosing an adaptive threshold for providing empathetic responses.

Embodied virtual agents and social robots that can emotionally engage their users have a huge potential in multiple domains including healthcare and education [11, 31]. With this work, we provide a blueprint for developing empathetic machines.

ACKNOWLEDGMENTS

We are thankful to Soheil Rayatdoost for his work and analysis on the speech spectrogram signals. This work was supported in part by the U.S. Army. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

REFERENCES

[1] F. Alam, M. Danieli, and G. Riccardi. 2018. Annotating and modeling empathy in spoken conversations. *Computer Speech & Language* 50 (2018), 40–61.

[2] T. Baltrusaitis, P. Robinson, and L. P. Morency. 2016. OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*. IEEE, 1–10.

[3] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 59–66.

[4] C. Becker, H. Prendinger, M. Ishizuka, and I. Wachsmuth. 2005. Evaluating affective feedback of the 3D agent max in a competitive cards game. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 466–473.

[5] T. W. Bickmore. 2003. *Relational agents: Effecting change through human-computer relationships*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[6] S. Brave, C. Nass, and K. Hutchinson. 2005. Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International journal of human-computer studies* 62, 2 (2005), 161–178.

[7] E. Cambria, I. Hupont, A. Hussain, E. Cerezo, and S. Baldassarri. 2011. Sentic avatar: Multimodal affective conversational agent with common sense. In *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*. Springer, 81–95.

[8] E. Cambria, B. Schuller, Y. Xia, and C. Havasi. 2013. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent systems* 28, 2 (2013), 15–21.

[9] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 163–171.

[10] C. Clavel and Z. Callejas. 2016. Sentiment analysis: from opinion mining to human-agent interaction. *IEEE Transactions on affective computing* 7, 1 (2016), 74–93.

[11] E. Deng, B. Mutlu, and M. J. Mataric. 2019. Embodiment in Socially Interactive Robots. *Foundations and Trends in Robotics* 7, 4 (2019), 251–356.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 248–255.

[13] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, et al. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1061–1068.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[15] P. Ekman and W. Friesen. 1978. *The Facial Action Coding System (FACS)*. Consulting Psychologists Press, Stanford University, Palo Alto.

[16] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* 7, 2 (apr 2016), 190–202.

[17] F. Eyben, F. Weninger, F. Gross, and B. Schuller. 2013. Recent developments in openSMILE, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia - MM '13*. ACM Press, New York, New York, USA, 835–838.

[18] F. Eyben, M. Wöllmer, and B. Schuller. 2010. OpenSMILE: The Munich Versatile and Fast Open-source Audio Feature Extractor. In *Proceedings of the 18th ACM International Conference on Multimedia (MM '10)*. ACM, New York, NY, USA, 1459–1462.

[19] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews.. In *LREC*. Citeseer, 3123–3128.

[20] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2122–2132.

[21] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[22] C. J. Hutto and E. Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

[23] I. Leite, A. Pereira, S. Mascarenhas, C. Martinho, R. Prada, and A. Paiva. 2013. The influence of empathy in human–robot relations. *International journal of human-computer studies* 71, 3 (2013), 250–260.

[24] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria. 2018. Dialoguernn: An attentive rnn for emotion detection in conversations. *arXiv preprint arXiv:1811.00405* (2018).

[25] A. Metallinou, A. Katsamanis, and S. Narayanan. 2013. Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing* 31, 2 (2013), 137–152.

[26] J. W. Pennebaker, M. E. Francis, and R. J. Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

[27] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 973–982.

[28] S. Poria, E. Cambria, and A. Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2539–2544.

[29] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* 174 (2016), 50–59.

[30] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain. 2016. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE, 439–448.

[31] B. Scassellati, J. Brawer, K. Tsui, S. Nasihati Gilani, M. Malzkuhn, B. Manini, A. Stone, G. Kartheiser, A. Merla, A. Shapiro, et al. 2018. Teaching language to deaf infants with a robot and a virtual human. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 553.

[32] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing* 65 (2017), 3–14.

[33] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan. 2010. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *Proc. INTERSPEECH 2010, Makuhari, Japan*. 2362–2365.

[34] B. Xiao, D. Can, P. G. Georgiou, D. Atkins, and S. S. Narayanan. 2012. Analyzing the language of therapist empathy in motivational interview based psychotherapy. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 1–4.

[35] B. Xiao, Z. E. Imel, P. G. Georgiou, D. C. Atkins, and S. 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), e0143055.

[36] H. Xiao. 2018. bert-as-service. https://github.com/hanxiao/bert-as-service.

[37] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* (2017).