Improving Assessments Using Intelligent Agents with Transient Emotional States

Angie Dowdell
Consortium Research Fellows Program
Fort Benning, GA
dowdell_angie@columbusstate.edu

Rania Hodhod, PhD
Columbus State University
Columbus, GA
hodhod_rania@columbusstate.edu

Suleyman Pölat
University of North Texas
Denton, TX
suleymanpolat@my.unt.edu

Randy Brou, PhD
Army Research Institute, Fort Benning
Fort Benning, GA
randy.j.brou.civ@mail.mil

Julia Grove
Consortium Research Fellows Program
Fort Benning, GA
jkw0034@auburn.edu

ABSTRACT

The assessment of interpersonal leadership skills has historically been impeded by the lack of assessment techniques free of response bias and resource prohibitions. Advances in psychology and computer science (e.g., Reactive, Open-Response Assessments; RORAs) have provided novel alternatives to traditional assessment methods (Brou, Stallings, Normand, Stearns, & Ledford, 2018). RORAs assess interpersonal skills via scenarios in which users interact with virtual agents; however, these agents currently lack sophisticated emotional response capabilities. The present research addresses this shortcoming by developing system capabilities to parse emotional inputs and generate appropriate emotional responses. Such techniques applied in other domains using intelligent agents have been shown to increase the effectiveness of information delivery and retrieval (McDuff & Czerwinski, 2018). Using Russell's Circumplex Model of Affect (Russell, 1980) as a theoretical framework, four models designed to classify the affect of RORA inputs and appropriately drive virtual agent behavior were developed using Google's DialogFlow Platform (Vibbert, et al., 2017). One-hundred and twenty-five participants provided conversational inputs intended to modify or maintain a virtual agent's affective states. When classifying both emotional valence and intensity of user inputs, model accuracy ranged between 46% and 58%. When classifying emotional valence alone (i.e., positive versus negative emotions), models correctly classified between 70% and 85% of user inputs. Next, deep learning techniques (e.g., sequence to sequence) were used to generate novel agent utterances based on current affective state. Novel agent utterances were evaluated by the researchers for intelligibility and appropriateness of emotional valence. Quality of the novel utterances was promising: 76% of the generated utterances were intelligible and, of those intelligible utterances, 91% exhibited appropriate emotional valence. These results demonstrate the potential for improving intelligent agent emotional response capabilities within an interpersonal skill assessment.

ABOUT THE AUTHORS

Angie Dowdell is a research fellow at the Army Research Institute and is a recent graduate student at the TSYS School of Computer Science, Columbus State University. As part of the Consortium Research Fellowship Program (CSRFP), she actively corresponds with a team that delivers technologies used for improving soldier training efficiency.

Rania Hodhod, PhD serves as the Assistant Chair and as an Assistant Professor in the TSYS School of Computer Science, Columbus State University. She teaches and supervises undergraduate and graduate students. Dr. Hodhod has published over 45 refereed articles within her specialized domain of research. This domain spans a range of areas, including artificial intelligence, expert systems, serious games, interactive narrative, and computational creativity.

Suleyman Pölat is a Graduate Assistant at the Department of Computer Science and Engineering at the University of North Texas. His work focuses on a user-centered data science approach to modeling and simulating virtual agents for communication training, as an assistant to Dr. Rodney Neilson. He has 4 years of experience as a software developer and data scientist.

Randy Brou, PhD is a Research Psychologist in the Army Research Institute, Fort Benning Research Unit. He has 20 years of experience in conducting applied research for the Department of Defense. His work has focused on training effectiveness and the measurement of individual and team attributes relevant for successful performance.

Julia Grove is a consortium research fellow at the Army Research Institute, Fort Benning unit. She is an industrial-organizational psychology Ph.D. student at Auburn University. Her research includes the use of gamified assessments to examine decision making processes.

Improving Assessments Using Intelligent Agents with Transient Emotional States

Angie Dowdell
Consortium Research Fellows Program
Fort Benning, GA
dowdell_angie@columbusstate.edu

Rania Hodhod, PhD
Columbus State University
Columbus, GA
hodhod_rania@columbusstate.edu

Suleyman Pölat
University of North Texas
Denton, TX
suleymanpolat@my.unt.edu

Randy Brou, PhD
Army Research Institute, Fort Benning
Fort Benning, GA
randy.j.brou.civ@mail.mil

Julia Grove
Consortium Research Fellows Program
Fort Benning, GA
jkw0034@auburn.edu

INTRODUCTION

In military and civilian organizations, leader effectiveness is critical to the attainment of organizational goals and overall organizational success. Both task and contextual performance impact leadership effectiveness (Motowidlo & Van Scotter, 1994). Interpersonal leadership skills are an important element of contextual performance (Mumford, Campion, & Morgeson, 2007; Van Scotter & Motowidlo, 1996) of particular interest to the US Army. In its publications on leadership, the Army places emphasis on interpersonal leadership skills such as empathy, interpersonal tact, and building trust (FM 6-22). Although the Army recognizes the vital importance of interpersonal leadership skills, the assessment and systematic development of those skills presents a challenge (Bedwell, Fiore, & Salas, 2014). Historically, the Army has implemented two traditional methodologies for the assessment of interpersonal leadership skills: self-report measures and live assessment. Self-report metrics demonstrate the advantage of ease of administration, but suffer from vulnerability to various forms of response bias (Donaldson & Grant-Vallone, 2002; Klehe, et al., 2012; Nederhof, 1985). Although live assessments demonstrate greater objectivity, these assessments are often prohibitively resource-intensive. The challenge of overcoming these weaknesses has prompted research through the years, and recent research has demonstrated the promise for a novel assessment method: reactive, open-response assessments (RORAs; Brou, Stallings, Normand, Stearns, & Ledford, 2018).

RORAs consist of a set of virtual agents (i.e., virtual human characters) and environments with which a respondent interacts via unguided, free-text responding. In a given assessment, a specific skill or attribute can be objectively assessed based on a respondent's behaviors in a tailored, interactive scenario. This assessment methodology combines the scalability and ease of use of self-report measures with the objective measuring of unguided responses possible in live assessments. In the initial work on RORAs, Brou et al. (2018) demonstrated that performance on the assessments significantly correlated with instructor assessments of interpersonal leadership skills in a junior Officer leadership course – a promising finding for the new methodology. However, a weakness in the RORAs was that agent behaviors were seen as unreasonable in about 20% of cases. One potential solution for making the agents seem more reasonable in their behavioral repertoire would be to develop more scripted reactions to be triggered under specific circumstances (i.e., adding more branches to the scenarios). This solution could quickly become problematic if one tries to imagine enough branches to account for an unknown breadth of possible participant inputs. Another solution is to develop ways to make agent behaviors more nuanced while still maintaining a manageable number of branches. Thus, the current effort explores the possibility of adding transient emotional state tracking and emotionally-informed, natural language generation to virtual agents like those used in the RORAs.

Representation of Emotions for Interpersonal Leadership Assessment

Previous research has demonstrated that interactions with software systems that mimic human-to-human interaction via the inclusion of capacity to respond to emotional cues lead to improved user engagement, trust, sense of rapport, etc. (McDuff & Czerwinski, 2018). Given that interpersonal leadership includes managing emotional expression in one's self and one's peers/subordinates, allowing for emotional exchanges that feel organic may enhance the ability of RORAs to assess relevant skills (Riggio & Lee, 2007). Koval and Kuppens (2012) define "inter-speaker emotional influence" as the degree to which one person's emotional expression influences another individual's emotional state. That is, emotional experiences are not entirely subjective, but vary as a function of socioenvironmental factors such as the emotional state of a conversational counterpart (Koval & Kuppens, 2012; Thornton & Tamir, 2017). If emotional dynamics tend to be mirrored between conversational counterparts, an important aspect of leadership would be demonstrating emotions consistent with the emotions one expects to elicit from one's subordinates. As the scope of prospective emotional responses to a spectrum of social interactions is enormous, the present research imposes a restriction of range in focus to a subset of emotional expressions relevant to interpersonal leadership skills: valence and intensity (Lang, 1995; Russell, 1980).

The Circumplex Model of Affect (CMA) was utilized in the present study for emotional classification due to its simplicity of structure and cross-cultural validity (Russell, 1980; Russell, Lewicka, & Niit, 1989). Figure 1 contains a visual depiction of the CMA, which includes four regions divided by two axes to classify emotional expression. The horizontal axis is the valence axis (i.e., the direction of an individual's emotional expression), with unpleasant emotion at one end of the continuum and pleasant emotion at the other. Unpleasant or negative emotions include feelings of nervousness, upset, depression, and boredom. Conversely, pleasant or positive emotions encompass feelings of elation, serenity, alertness, and comfort (Russell, 1980). The vertical axis is the intensity axis, which denotes the degree of emotional intensity expressed by an individual. High intensity emotional expression includes emotions such as frustration, happiness, stress, and excitement, while low intensity emotional expression includes emotions such as complacency, lethargy, calmness, and fatigue (Russell, 1980). These two perpendicularly intersecting axes delineate four distinct regions: Region 1 consists of positive valence and high intensity (PVHI) emotions such as happiness, enthusiasm, and excitement; Region 2 includes negative valence and high intensity (NVHI) emotions such as frustration, stress, and nervousness; Region 3 is comprised of negative valence and low intensity (NVLI) emotions such as sadness, depression, and lethargy, and Region 4 encompasses positive valence and low intensity (PVLI) emotions such as complacency, content, and calm. For the purposes of the current research, Region 1 will henceforth be referred to as PVHI, Region 2 as NVHI, Region 3 as NVLI, and Region 4 as PVLI.

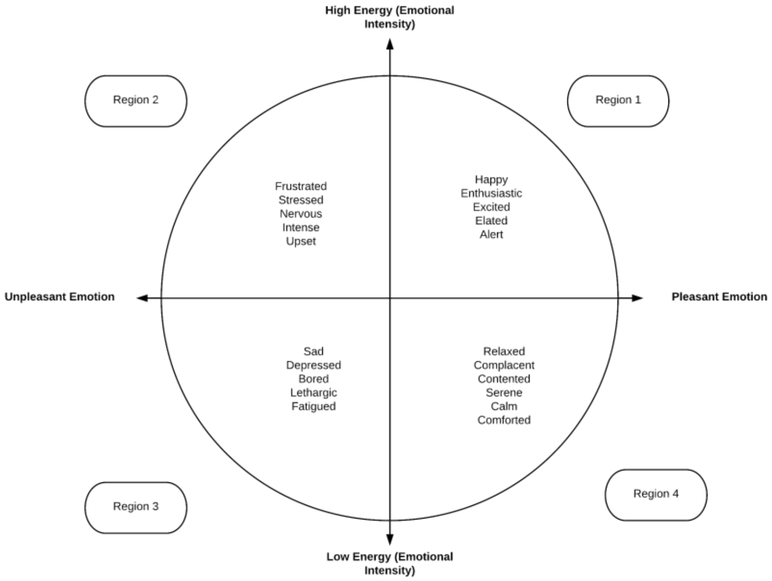


Figure 1. Circumplex Model of Affect (Russell, 1980)

Emotional Input and Output

Developing agents able to react as a function of emotional states requires two separate considerations: affect sensing and natural language generation. Affect sensing is the ability to interpret the emotional content of inputs, while natural language generation is the ability to produce intelligible utterances given a conversational context (e.g., emotional states). One of the most popular affect sensing techniques is keyword spotting. This technique categorizes utterances into affective classifications using words with obvious emotional salience (e.g., happy, despondent, frustrated; Elliott, 1989). While this technique is popular and relatively easy to carry out, results may be suboptimal due to the high emphasis on surface-level feature extraction. For example, the sentence "Today is a new day for me," may yield scant affective information computationally, but may be high in emotional intensity. Statistical natural language processing (e.g., support vector machine) techniques work well for large text parsing but yield poor results for sentence or word level parsing (Elliott, 1989).

Lui, Lieberman, and Selker (2003) explored a novel approach to textual affect sensing using a real-world, common sense knowledge base. A corpus of nearly half a million sentences was used to evaluate the affective nature and underlying semantic structure of presented sentences. Sentences were emotionally categorized using the Ekman emotional model (Ekman, 1999). The approach went beyond surface-level feature extraction techniques seen in traditional approaches to textual affect sensing, analyzing underlying semantic meanings related to affect on a sentence level. The approach could not, however, account for higher-level conversational contexts – a feature desirable when driving the behavior of agents through a sequence of interpersonal interactions.

In order to accomplish the goal of accounting for conversational context, the present research sought to employ a combination of Lui, et al's (2003) approach with recent developments in sequence to sequence (seq2seq) dialog modeling (Sutskever, Vinyals, & Le, 2014). Seq2seq refers to a machine learning architecture that can vectorize language inputs in a way that takes linguistic ordering into account. This means that the algorithms involved don't just process individual words for their emotional salience, but rather they incorporate contextual cues provided by surrounding words. Contextual cues can be gathered from more than just the immediate sentence being parsed. Seq2seq models can learn from a contextually appropriate dialogue corpus developed prior to a given interaction. Moreover, through such models, the entire history of a given interaction between an agent and human user could become relevant for understanding the emotional salience of utterances, giving agents a kind of emotional memory through which behaviors could be driven. For the present research, a dialogue corpus was generated to support affect sensing and natural language dialogue generation applicable to social interactions within a military context. In the following sections, the development of the dialogue corpus, its instantiation into systems for affect sensing and natural langue generation, and the performance of the resulting models will be discussed.

METHOD

Overview

To facilitate the goals of agent affect sensing and natural language generation, a tailored dialogue corpus was needed for model training. The Real World Professional Conflicts Dialogue Corpus (RWPCDC) was developed to reflect the four emotional regions (PVHI, NVHI, NVLI, and PVLI) on the CMA. The RWPCDC was then used as the basis for training and testing models for both affect sensing and natural language generation. Affect sensing performance was evaluated by measuring the accuracy with which models could classify unfamiliar emotional statements, while natural language generation performance was evaluated in terms of utterance intelligibility and emotional appropriateness.

Building the Real World Professional Conflicts Dialogue Corpus

Participants

Two groups of participants contributed to the building of the RWPCDC. The first group consisted of 96 students from a junior Officer leadership course at a large Army training installation. The qualifications to enroll in the course include: being a US citizen, having a 4-year college degree, and being between 19 and 32 years old. No specific demographics were collected for the students, but most (~80%) would have been males in their early 20s. The second group consisted of 29 participants from psychology department of a public university. No specific demographics were collected for this group either, but most (~59%) would have been females between 18 and 24 years old.

Materials and Procedures

Participants provided utterances in response to four text-based scenarios (each based on existing Army training materials), which are presented in Table 1. Each scenario presented a situation in which a character is introduced with whom the participant is engaged in conversation. Participants took on the role of a leader (i.e., Lieutenant) who is conversing with a professional subordinate in the context of a work-based conflict. The subordinate characters in the four scenarios vary in their initial emotional state. That is, one character is depicted as PVHI (positive in valence, high in intensity – happy, enthusiastic), another as NVHI, etc. (Russell, 1980). For each scenario, participants were asked to provide eight utterances (each with a 255-character limit). Six of these utterances were to have the emotional impact of moving the character from their initial state to a specified target state (e.g., two utterances to move the character from PVHI to NVHI; two utterances to move the character from PVHI to NVLI; and two utterances to move the character from PVHI to PVLI). The final two utterances were to have the emotional impact of keeping the character in their initial state (e.g., to keep a character initially expressing a PVHI affective state in the PVHI region of the CMA; Russell, 1980). Thus, each participant generated 32 utterances in total across the four scenarios representing all the possible emotional transitions within the CMA.

Utterances from the junior Officer leadership course students were collected during proctored sessions, while the university student sample's utterances were collected via an online survey tool (Qualtrics). University participants received varying degrees of extra credit as determined by instructors in the psychology department, while the junior Officer leadership course students were not specifically incentivized. To optimize the quality of the dataset, utterances that were instructional as opposed to conversational (e.g., "Tell the soldier that he needs to straighten up his attitude or he's out") were omitted. The final RWPCDC dataset of over 3,400 utterances was utilized in the training of both the affect sensing and the natural language generation models.

Table 1. Scenarios by Circumplex Model Region

CMA Region	Scenario Description
PVHI (Region 1)	*"You are 2LT, 3rd Platoon, "B" Company, 1st Bn, 66th Infantry. You are introducing yourself to SFC Johnson as the new platoon leader. You inform him that you've heard about the outstanding work he has done in the recent absence of a platoon leader and express your excitement about serving with him. He is clearly elated to hear the news of this positive review.* *SFC Johnson explains that the company is in good shape as far as personnel is concerned the first and second squads are full strength, but the third and fourth squad are each one man short. Through this conversation, it becomes apparent that the squad leaders, team leaders, and NCOs are hardworking company members who get the job done. SFC Johnson continues to further elaborate upon the company weaknesses, mentioning that the 2nd Squad leader, Sgt. Cramer is an avid drinker who has the tendency of harassing the other squad leaders every once in a while. He ensures you that while this has been an issue he periodically pulls Sgt. Cramer to the side to set him straight, when necessary."*
NVHI (Region 2)	*"While walking down the sidewalk, you hear yelling and profanity coming from around the corner. You turn the corner and see SSG Burch (an E-6 Squad Leader) and his squad. SSG Burch is verbally abusing one of the soldiers in front of the entire squad.* *SSG Burch: 'Give me that weapon, Jones (snatching the rifle from Jones), you dumb idiot! How many times do I have to tell you to clean this weapon? Did your parents ever have any children that weren't retarded?'* *Overhearing, you call SSG Burch over and ask him to walk with you to your office. While in your office, you express your concerns about his approach, informing him that verbal abuse is not the way to get the job done. He responds saying, 'Yes, Sir but it's hard when the troops are always screwing up - and I've been under a lot of stress lately.' He reveals that he has been experiencing a high level of family-related stress due to a few disagreements that have occurred with his wife. He continues, saying, 'Last week, I went to Happy Hour and didn't get home 'till after ten o'clock. We had a hell of an argument and I wound up sleeping on the couch for three nights...' He explains that the arguments with his wife are starting to become more frequent, '...either it's about the kids, or the bills, or my mother-in-law, or the car. It's just always something!'"*

NVLI (Region 3)	*"You have been a Platoon Leader in B Company for 5 weeks. Today you are counseling PFC Lewis, who is has been among the best men in your outfit. However, in the past few weeks, he has noticeably demonstrated abrupt changes in behaviors. Recognizing that these changes could be indicative of more serious issues, you began to inquire about recent events asking, 'PFC Lewis, I understand that over the last couple of weeks your performance has changed. SGT Franklin tells me you've been late to morning formation three times in the last two weeks. What's been going wrong?' Despondently, he responds revealing that in the past few weeks the recent divorce of his parents and the split from his fiancé of 2 years has begun to take a toll on him. He continues saying, 'Sir, these last few weeks have been terrible. Sometimes I really get down. A few times I have even thought of ending it all for good… One of these times I'm afraid I might really…do it.'"*
PVLI (Region 4)	*"You have been the Mortar Plt. Ldr. in A Company for the past two months. During this time you have had numerous discussions with SFC Smothers, your Plt Sgt and have determined that your sections leaders are highly motivated and proficient at their jobs. However, you have observed that SSG Roger's the first Section leader has erratic performance. Two weeks ago SSG Roger's squad was detailed to perform police call at 0700 hours around the post HQ. He did not get his men there until 0830. During a counseling session with SFC Smothers, SSG Rogers said he failed to arrange for transportation in advance. Yesterday afternoon, one of his vehicles failed a roadside spot inspection. You are scheduled to meet SGT Rogers at 0800 this morning, in order to discuss the recent events. He is running late and does not report to the counseling session until 0815. Just outside of your office you hear him making small talk with an NCO, saying that he had turkey sausages and salsa scrambled eggs, his favorite meal for breakfast, as he slowly makes his way to your door. He walks in and greets you calmly saying 'Sir, Staff Sergeant Rogers reports.' You usher him to come in."*

Training and Testing Models

Affect Sensing

Google's DialogFlow was used for the training and evaluation of four affect sensing models corresponding to the four scenarios for which utterances were collected (Vibbert, Goussard, Beaufort, & Monnahan, 2017). Participants were asked to provide utterances that would either shift the emotional state of the model to a different region of the CMA or maintain the current emotional state in a specified region of the CMA. The models each began in a different region of the CMA as indicated by their labels (e.g., PVHI denotes the model that began with positive valence and high intensity). When provided with a user utterance, the model was expected to distinguish the emotional intent of the utterance. This would then establish the conditions for appropriately modifying the emotional state of the character/agent within a RORA. Over 800 utterances from the RWPCDC were utilized for the training and testing of each model. Within the models, a machine learning classification threshold was set to a confidence score of 0.3, meaning that inputted utterances corresponding with classification confidence scores less than 30% were left uncategorized. Training data (90% of the RWPCDC) was first supplied to each model. Model testing was then conducted by submitting the reserved 10% of utterances and determining whether the models' classifications matched the intent (i.e., the impact on emotional state) of the utterances supplied by the participants.

Natural Language Generation

Following the evaluation of the models' affect sensing accuracy, one model (NVHI) was chosen for natural language generation testing. In addition to the RWPCDC, the model was provided with a set of researcher-generated responses to specific RWPCDC inputs. For example, the RWPCDC utterance, "I don't care what is happening in your personal life; sort it out and don't verbally harass your Soldiers!" was paired with the response, "The only one guilty of verbal harassment around here is you, Sir!" In this way, the model was provided a template for the generation of novel, emotionally appropriate utterances. Model training was supplemented using the Cornell Movie Dialogue Corpus (Danescu-Niculescu-Mizil & Lee, 2011). This corpus provided a large set of conversational exchanges (over 220,000) to help the model learn to form intelligible sentence structures. Following model training, testing was conducted by providing the NVHI model with 30 unfamiliar utterances from the RWPCDC that fell within the NVHI region of the CMA. The goal for the model was to generate intelligible and emotionally appropriate (i.e., angry) responses to each of these utterances. Human raters evaluated each generated utterance for intelligibility as well as for the presence of the intended affect valence.

RESULTS AND DISCUSSION

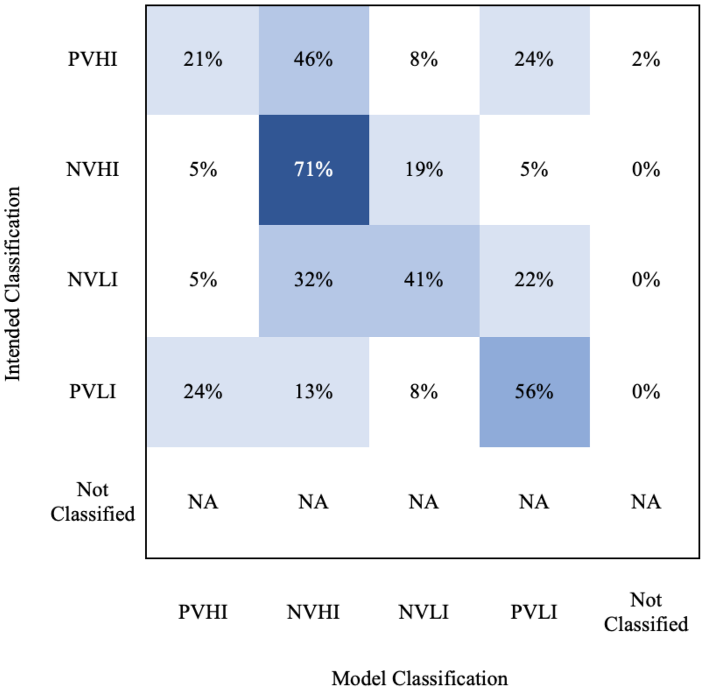
Affect Sensing

Accuracy rates for affect sensing among the four models are presented in Table 2. Overall accuracy refers to classification of an utterance into the correct quadrant of the CMA (e.g., if the utterance was intended to express anger, the utterance would be accurately classified as NVHI). Valence accuracy refers to classification of utterances into the correct half of the horizontal valence axis of the CMA (e.g., an utterance that was intended to express a positive emotion would achieve valence accuracy if it is classified as PVHI or PVLI). Intensity accuracy refers to classification of utterances into the correct half of the vertical intensity axis of the CMA (e.g., an utterance that was intended to express high emotional intensity would achieve intensity accuracy if it is classified as PVHI or NVHI).

Table 2. CMA Classification Accuracy per Model

Model	Overall Accuracy	Valence Accuracy	Intensity Accuracy
PVHI (Region 1)	46.48%	70.70%	66.41%
NVHI (Region 2)	58.14%	84.88%	67.44%
NVLI (Region 3)	48.75%	77.50%	57.50%
PVLI (Region 4)	51.25%	81.25%	58.75%

Figures 2, 3, 4, and 5 are confusion matrices for model accuracy, which depict the percentage of emotional classifications for each type of input provided to the model. In the confusion matrices, "intended classification" refers to the region of the CMA for which the utterance was crafted by participants (e.g., an intended classification of PVHI would accurately be classified as PVHI). "Model classification" refers to the region of the CMA within which the model's affect sensing algorithm categorized the utterance (e.g., an utterance was classified within PVHI, regardless of the participants' intent). Correct classification occurs when the model classification region is congruent with the intended classification region.

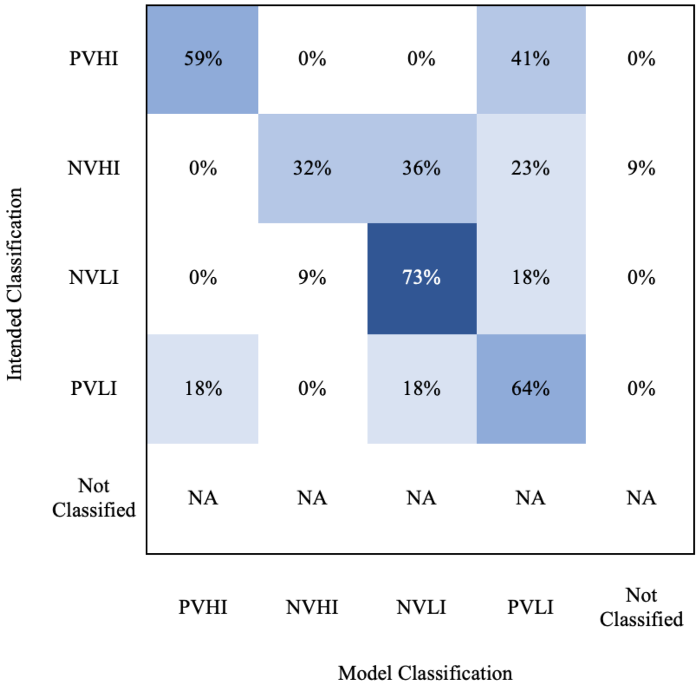


Intended Classification \ Model Classification	PVHI	NVHI	NVLI	PVLI	Not Classified
PVHI	21%	46%	8%	24%	2%
NVHI	5%	71%	19%	5%	0%
NVLI	5%	32%	41%	22%	0%
PVLI	24%	13%	8%	56%	0%
Not Classified	NA	NA	NA	NA	NA

Figure 2. Confusion Matrix, PVHI (Region 1) Performance

Figure 2 depicts the accuracy of affect sensing for the PVHI model. The model correctly classified utterances as PVHI 21% of the time, correctly classified utterances as NVHI 71% of the time, correctly classified utterances as NVLI 41%

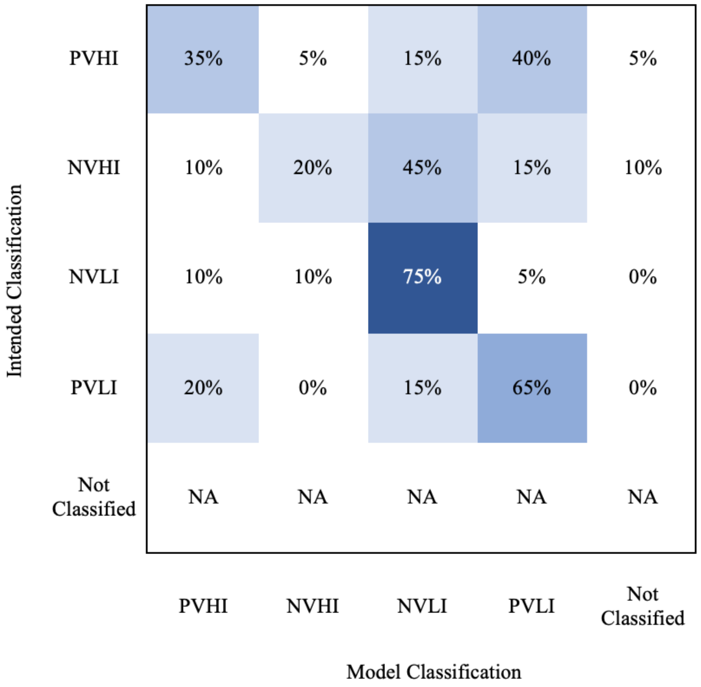
of the time, and correctly classified utterances as PVLI 56% of the time. The model was most likely to misclassify utterances intended to evoke a PVHI state as NVHI, which occurred 46% of the time. The model was most likely to misclassify utterances intended to evoke a NVHI state as NVLI, which occurred 19% of the time. Utterances intended to evoke a NVLI affective state were most likely to be misclassified as NVHI, which occurred 32% of the time. The model was most likely to misclassify utterances intended to evoke a PVLI affective state as PVHI, an error that occurred 24% of the time. Misclassifications in the NVHI, NVLI, and PVLI regions of the CMA generally represent issues with the capacity of the model to detect emotional intensity, while the misclassification trend for the PVHI region represents an issue with the capacity of the model to detect emotional valence.



	PVHI	NVHI	NVLI	PVLI	Not Classified
PVHI	59%	0%	0%	41%	0%
NVHI	0%	32%	36%	23%	9%
NVLI	0%	9%	73%	18%	0%
PVLI	18%	0%	18%	64%	0%
Not Classified	NA	NA	NA	NA	NA

Figure 3. Confusion Matrix, NVHI (Region 2) Performance

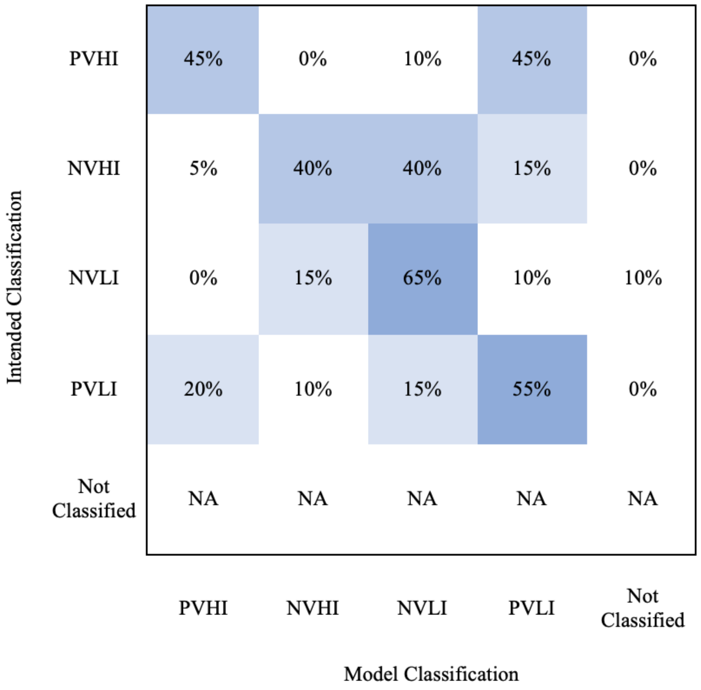
Figure 3 demonstrates the accuracy of affect sensing for the NVHI model. This model correctly classified utterances as PVHI 59% of the time, correctly classified utterances as NVHI 32% of the time, correctly classified utterances as NVLI 73% of the time, and correctly classified utterances as PVLI 64% of the time. The model was most likely to misclassify utterances intended to evoke a PVHI affective state as PVLI, which occurred 41% of the time. The model was most likely to misclassify utterances intended to evoke a NVHI state as NVLI, which occurred 36% of the time. Utterances intended to evoke a NVLI affective state were most likely to be misclassified as PVLI, which occurred 18% of the time. The model was most likely to misclassify utterances intended to evoke a PVLI affective state as either PVHI, an error that occurred 18% of the time, or as NVLI, an error that occurred 18% of the time. Misclassifications in the PVHI and NVLI regions of the CMA represent issues with the capacity of the model to detect emotional intensity. The misclassification trend for the NVLI region represents an issue with the ability of the model to detect emotional valence, while the misclassification trend for the PVLI region represents an issue with the capacity of the model to detect emotional valence and emotional intensity.



	PVHI	NVHI	NVLI	PVLI	Not Classified
PVHI	35%	5%	15%	40%	5%
NVHI	10%	20%	45%	15%	10%
NVLI	10%	10%	75%	5%	0%
PVLI	20%	0%	15%	65%	0%
Not Classified	NA	NA	NA	NA	NA

Intended Classification
Model Classification

Figure 4. Confusion Matrix, NVLI (Region 3) Performance



	PVHI	NVHI	NVLI	PVLI	Not Classified
PVHI	45%	0%	10%	45%	0%
NVHI	5%	40%	40%	15%	0%
NVLI	0%	15%	65%	10%	10%
PVLI	20%	10%	15%	55%	0%
Not Classified	NA	NA	NA	NA	NA

Intended Classification
Model Classification

Figure 5. Confusion Matrix, PVLI (Region 4) Performance

Figure 4 depicts the accuracy of the affect sensing for the NVHI model. This model correctly classified utterances as PVHI 35% of the time, correctly classified utterances as NVHI 20% of the time, correctly classified utterances as NVLI 75% of the time, and correctly classified utterances as PVLI 65% of the time. The model was most likely to misclassify utterances intended to evoke a PVHI state as PVLI, which occurred 40% of the time. The model was most likely to misclassify utterances intended to evoke a NVHI state as NVLI, which occurred 45% of the time. Utterances intended to evoke a NVLI affective state were most likely to be misclassified as either PVHI or as NVHI, both of which occurred 10% of the time. The model was most likely to misclassify utterances intended to evoke a PVLI affective state as PVHI, an error that occurred 20% of the time. Misclassifications in the PVHI, NVLI and PVLI regions of the CMA represent issues with the capacity of the model to detect emotional intensity. The misclassification trend for the NVLI region represents an issue with the capacity of the model to detect emotional valence and intensity.

Figure 5 depicts the accuracy of affect sensing for the PVLI model. This model correctly classified utterances as PVHI 45% of the time, correctly classified utterances as NVHI 40% of the time, correctly classified utterances as NVLI 65% of the time, and correctly classified utterances as PVLI 55% of the time. The model was most likely to misclassify utterances intended to evoke a PVHI state as PVLI, which occurred 45% of the time. The model was most likely to misclassify utterances intended to evoke a NVHI state as NVLI, which occurred 40% of the time. Utterances intended to evoke a NVLI affective state were most likely to be misclassified as NVHI, which occurred 15% of the time. The model was most likely to misclassify utterances intended to evoke a PVLI affective state as PVHI, an error that occurred 20% of the time. Misclassifications in all regions of the CMA represent issues with the capacity of the model to detect emotional intensity.

As indicated by the results of the affect sensing task, accuracy rates for valence classification were better than accuracy rates of intensity classification. This is likely due to the mode of communication, as intensity of emotions is not as readily interpretable via written communication compared to face-to-face communication. That is, individuals differentiate between utterances intended to express positive versus negative emotions by changing the substantive content of the utterance (Strapparava, et al., 2006; Valitutti, et al., 2004). For instance, to express a positive emotion, an individual may provide input such as "Today was a good day," while negative emotional expression may consist of utterances such as "Today was a bad day." On the other hand, expressions of emotional intensity may vary as a function of tone of voice or body language, which are not as readily detected by classification algorithms (Strapparava, et al., 2006; Valitutti, et al., 2004). For instance, an individual may express a high intensity emotion with the phrase "Today was a good day!!" but express a lower-intensity iteration of the same emotion with the phrase "Today was a good day." Furthermore, the distinction between the emotional intensity of utterances may be made using subtle, rather than drastic, changes in terminology. For instance, a high-intensity positive expression might be "Today was a great day," while a low-intensity positive expression might be "Today was a good day." As such, the difference between a high intensity emotional expression and a low intensity emotional expression is indicated via punctuation, not by differences in the lexical content of utterances. Therefore, emotional valence may be more readily detected than emotional intensity.

Given the results of the affect sensing portion of the study, the natural language generation portion of the present study focused on the generation of utterances congruent with intended emotional valence without specific regard to intensity.

Natural Language Generation

The results of the natural language generation task indicated that 23/30 (76.67%) of the machine-generated utterances were intelligible. Out of the 23 intelligible responses, 21 (91.30%) of the responses matched the intended affective valence (e.g., negative emotional valence). Table 3 contains examples of intelligible, unintelligible, intended valence, and unintended valence utterances as well as the percentage of total utterances represented by each category.

Table 3. Natural Language Generation Samples (Percentage of Generated Responses)

	Intelligible	**Not Intelligible**
Intended affect valence	"Wait a minute, don't tell me how to do my job" (66.67%)	"You cannot! You are a sorry excuse for a treat you to fight" (16.67%)
Unintended affect valence	"Yes sir I will, I am a professional" (10.00%)	"It is a direct reflection on it" (6.67%)

The performance of the model on the natural language generation task was quite promising. The majority of the utterances were both intelligible and possessed the appropriate affect valence. The model's most significant problem was producing fully intelligible utterances. Even for utterances deemed unintelligible, it is apparent that sections of the utterances made sense. However, the sensible portions of generated utterances did not always fit together properly, or were not paired with other portions necessary for completing a thought.

Limitations

The current research demonstrated the potential for using affect sensing and natural language generation to improve agent responses during interpersonal assessments. Issues remain for sensing affect intensity and generating reliably intelligible utterances. Larger training samples may improve performance, as may using a more appropriate general conversational corpus. The general corpus (Cornell Movie Dialogue Corpus) used for the current research may not have been ideal for the intended purposes given many of its themes are contextually inappropriate (e.g., romance, murder, and other movie-relevant tropes). Another general limitation was participant confusion during data collection. Participant confusion is evidenced by several responses taking the form of general descriptions of actions rather than the conversational utterances requested in the directions (e.g., "to make him feel like they want his help"). Differences in user responses can be normalized in the future by providing additional instructional guidance.

CONCLUSION

Two conversational modeling tasks were explored during the current research: 1) an affect sensing task and 2) a natural language generation task. The models for the first task categorized the emotional valence and emotional intensity of given utterances within a confined representation of emotional states. The predictive model for the second task generated novel expressions that corresponded with utterances expected given predefined conversational contexts. The models' ability to identify and respond to utterances as a function of emotional valence appears to operate at a satisfactory level of accuracy. Additional work is needed to account for emotional intensity. This may include moving towards a voice-recognition-based input system as opposed to a purely text-based system to account for tone of voice data. An encouraging aspect of these findings is that model performance was reasonable despite what would traditionally be considered a small sample size. Many seq2seq models are trained and evaluated using thousands or millions of training utterances, while the present model utilized only about 200 utterances per CMA region per model. This seems to indicate that a functional model could be produced with only a moderate amount of investment. Future work will need to be conducted to integrate the natural language advances made in this research with those of the existing RORA systems to evaluate the overall impact of modeling transient emotional states on interpersonal leadership skill assessment.

REFERENCES

Bedwell, W. L., Fiore, S. M., & Salas, E. (2014). Developing the future workforce: An approach for integrating interpersonal skills into the MBA classroom. *Academy of Management Learning and Education*, 13(2), 171-186.

Brou, R., Stallings, G., Stearns, I., Normand, S., & Ledford, B. (2018). Building automated assessments of interpersonal leadership skills. Paper presented at the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, FL.

Danescu-Niculescu-Mizil, C., & Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the workshop on cognitive modeling and computational linguistics*. ACL.

Donaldson, S. I. & Grant-Vallone, E. J. (2002). Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology*, 17(2), 245-262.

Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, 45-60.

Elliott, C. (1989). *The Affective Reasoner: A Process Model of Emotions in a Multiagent System* (Doctoral dissertation, PhD thesis, Northwestern University, May 1992. The Institute for the Learning Sciences, Technical Report).

Klehe, U. C., Kleinmann, M., Hartstein, T., Melchers, K. G., Konig, C. J., Heslin, P. A., & Lievens, F. (2012). Responding to personality tests in a selection context: The role of the ability to identify criteria and the idea employee factor. *Human Performance*, 25(4), 273-302.

Koval, P., & Kuppens, P. (2012). Changing emotion dynamics: individual differences in the effect of anticipatory social stress on emotional inertia. *Emotion*, *12*(2), 256.

Lang, P. J. (1995). The emotion probe: studies of motivation and attention. *American psychologist*, *50*(5), 372.

Liu, H., Lieberman, H., & Selker, T. (2003, January). A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th international conference on Intelligent user interfaces* (pp. 125-132). ACM.

McDuff, D., & Czerwinski, M. (2018). Designing emotionally sentient agents. *Communications of the ACM*, *61*(12), 74-83.

Motowidlo, S. J., & Van Scotter, J. R. (1994). Evidence that task performance should be distinguished from contextual performance. *Journal of Applied psychology*, *79*(4), 475.

Mumford, T. V., Campion, M. A., & Morgeson, F. P. (2007). The leadership skills strataplex: Leadership skill requirements across organizational level. *The Leadership Quarterly*, 18, 154-166.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology,* 15(3), 263-280.

Riggio, R.E., & Lee, J. (2007). Emotional and interpersonal competencies and leader development. *Human Resource Management Review*, 17 (4), 418-426.

Robinson, P., & Baltrušaitis, T. (2015, September). Empirical analysis of continuous affect. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 288-294). IEEE.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Russell, J. A., Lewicka, M., & Niit, T. (1989). A cross-cultural study of a circumplex model of affect. *Journal of personality and social psychology*, *57*(5), 848.

Sutskever, I., Vinyals, O., & Le, Q.V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104-3112).

Strapparava, C., Valitutti, A., & Stock, O. (2006, May). The Affective Weight of Lexicon. In *LREC* (pp. 423-426).

Thornton, M. A., & Tamir, D. I. (2017). Mental models accurately predict emotion transitions. *Proceedings of the National Academy of Sciences*, *114*(23), 5982-5987.

Valitutti, A., Strapparava, C., & Stock, O. (2004). Developing affective lexical resources. *PsychNology Journal*, *2*(1), 61-83.

Van Scotter, J. R., & Motowidlo, S. J. (1996). Interpersonal facilitation and job dedication as separate facets of contextual performance. *Journal of applied psychology*, *81*(5), 525.

Vibbert, M., Goussard, J. O., Beaufort, R. J., & Monnahan, B. P. (2017). *U.S. Patent No. 9,767,794*. Washington, DC: U.S. Patent and Trademark Office.

The research described herein was sponsored by the Army Research Institute for the Behavioral and Social Sciences, Department of the Army. The views expressed in this paper are those of the author and do not reflect the official policy or position of the Department of the Army, DOD, or the U.S. Government.