3-2019

Conversational Agent: Developing a Model for Intelligent Agents with Transient Emotional States

Angie Dowdell

CONVERSATIONAL AGENT: DEVELOPING A MODEL FOR INTELLIGENT AGENTS WITH TRANSIENT EMOTIONAL STATES

Angie Dowdell
2019

Columbus State University

D. Abbott Turner College of Business and Computer Science

The Graduate Program in Applied Computer Science

Conversational Agent: Developing a Model for Intelligent Agents with Transient Emotional States

A Thesis in
Applied Computer Science
by

Angie Dowdell

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

March 2019

Contents

List of Figures

ACKNOWLEDGEMENTS

First and foremost, I am deeply grateful to God for guidance through this process. I would like to thank my advisers for their continual support and diligence, Dr. Rania Hodhod and Dr. Randy Brou, as well as my committee members, Dr. Shamim Khan and Dr. Shuangbao (Paul) Wang. I am greatly appreciative of my family and friends for their support and encouragement.

Chapter 1

Introduction

The demonstration of aggressive behaviors in the workplace can usually lead to negative professional consequences. However, combat-related professions may involve life and death decision making on an almost daily basis. This is one occupational area in which major duties may invoke or even require aggression. Today's Army leaders face a different set of challenges than those faced by leaders of the past. Leaders of today must be trained to be both assertive and aggressive at the appropriate times while also displaying empathy. While this is not a new task for an Army leader, it is one that is approached differently among millennials.

Millennials are individuals who became adults in the early 21st century [2]. These individuals may have adapted to a communication style that is not as directly confrontational as the style seen in the former generation of Army leaders. This may be due to the heavy use of technology-oriented communication channels (e.g., texting, social media), which often occur in the place of face to face styles of communication. Army Basic Officer Leadership instructors face the challenge of preparing young leaders for combat and also guiding them into using more confrontational communication styles while exhibiting high levels of empathy. Such characteristics may prepare young officers to be decisive and effective leaders during combat.[2]

One of the most important aspects of those Army leadership skills is the interpersonal skill-set. Army leaders must be able to understand and value the perspectives of their comrades while balancing this understanding with their own perspectives enough to make clear and effective team-based decisions. On top of all of this, they must also be able to make these decisions quickly while under-pressure and, often-times, while physically depleted. The ability to carry out emotionally conscientious conversation can be one of the most imperative life skills to possess. For Army leaders who are working in a high-intensity environment in which many decisions can affect and influence an entire squad of subordinates, in such cases, the ability to carry out effective, emotionally intelligent dialogue within a short time frame can be a matter of life and death or toxic leadership and effective leadership. Thus, instructors who are training officers focus heavily on determining what kinds of interpersonal leadership attributes junior officers have and how to develop them. To develop such interpersonal skills instructors and researchers must first assess the initial interpersonal skill-set of student officers.[2]

1.1 Weaknesses of Traditional Assessments

The most commonly used assessments for interpersonal skills are traditional assessments. A traditional assessment is one that uses forced-choice measurement items (e.g., multiple choice, true and false, fill-in-the-blank, matching responses). Such assessments mainly evaluate a student's ability to recall previously obtained information and do not necessarily require demonstration or higher order application of that knowledge [3]. The issue with traditional assessments is that they do not always provide an accurate picture of how much a student may know or whether that knowledge can be applied effectively within a real-world context. This limitation may present a problem since such assessments may yield misleading results and leave students ill-prepared for handling real-world problems (Schwartz and Arena, 2013). Despite their weaknesses, traditional assessments are still highly popular because they are cheap, relatively easy to construct, produce quantifiable results, and are easy to administer [3].

In order to remedy the weaknesses found among traditional assessments, researchers like, Schwartz, are proposing that performance assessments become an alternative to traditional assessments. A performance assessment is an assessment that requires students to formulate solutions rather than choose from a limited sample of proposed solutions and to demonstrate abilities in a given area (e.g., medical school students' clinical skills assessments, driving tests, Army Physical Fitness Tests, writing assessments). Performance tests can be costly and are not as quick and easy to construct and administer as traditional assessments. One way to create a cost and time effective performance test is through the development of a model for interactive intelligent conversational agents with transient emotional states that can be used for assessment purposes. Through interacting with those intelligent agents, student officers may be able to demonstrate their interpersonal skill sets (empathy and perspective taking). [3]

1.2 Thesis Statement

The long-term aim of this research is to develop a computational model for an intelligent conversational agent with transient emotional states, in order to potentially support future research aims at the Army Research Institute, at Fort Benning. Such an agent may be embedded within a simulation environment that examines junior officers' empathy and perspective-taking as attributes of leadership. The junior officers in OCS courses may be able to demonstrate these skills through their responses to utterances within interpersonal Army training scenarios, in future phases of research. For the current research scope for this thesis project, the success of the model is determined by how well the model maps utterances to four regions on the Circumplex Model of Affect [4] based on their intended or expected emotional impact. This outcome will be assessed, in both a quantitative and qualitative manner.

Future ARI research aims that the current research thesis research may leverage or support include the following:

1. "Allowing nonlinear conversations to unfold" [5]
2. "Making agents more flexible by tracking emotional states, etc. across scenarios" [5]
3. "Identify scenario characteristics most responsible for improved language matching and better predictive validity to improve assessment techniques overall"[5]

The aim of the current research is to develop a generalized framework that will allow for the development of a generative affective conversational model that will be used to as part of an objective assessment of interpersonal Army leadership traits. Thus the goals of the current research aims are the following:

1. Develop a Corpus of Professional and Army Based Dialogue for the purpose of supporting the development of a conversational model

2. Develop a model that receives open-ended text input, computationally evaluates the affective quality of the input and provides a contextually appropriate outputted response

3. Provide a general framework for a conversational model that generates human-level affective conversation between a human user and intelligent agent (i.e., chatbot) in a nonlinear, generative manner.

4. Qualitatively assess the differences in the classification-based chatbot performance; identify nuances in the datasets and differences in the presentation of scenarios

1.3 Problem Statement and Our Contribution

One pertinent problem that will be addressed using these techniques in the current research is the assessment of interpersonal leadership skills of Army leaders through the use of conversational agents. As previously mentioned, traditional assessments are plagued with a plethora of weaknesses, recent developments like deep learning-based intelligent systems can facilitate a more well-rounded structure for identifying communication patterns of subjects. Additionally, deep learning approaches allow for the generation of helpful feedback in order to gauge and increase the effectiveness of learning, in an objective manner.

The current research will mainly explore the application of the deep learning algorithm to the development of a generalized framework that will be utilized in future research at the Army research institute at MCoE (Maneuver Center Of Excellence) facilities. The aim of ongoing research is to train junior officers in basic leadership courses. The algorithmic python-based implementation of the generative conversation models is adapted from two primary code sources. These code sources are inspired by the following Udemy Tutorial:*Deep Learning and NLP A-Z^{TM}- How to create a ChatBot.* The source code is documented and cited in Appendix D. For the purpose of current research, several customized modifications will be made to the original code source in order to support the current research aim [6]. This generalized framework will mainly be based upon modeling empathetic conversation between an intelligent conversational agent (chatbot/bot) and a human user (e.g., junior officer student in Army leadership courses) in order to allow for higher level observation of communication patterns. This generalized framework will contribute to the long-term and continued efforts of overall improvements in the training of interpersonally effective Army leaders.

1.4 Thesis Organization

This thesis is organized as follows: Chapter 2 provides a background of the research aim and a literature review of the related works. This chapter also presents the foundational framework that supports the current investigation.

Chapter 3 provides an overview of conversational agents as well as recent approaches to modeling interaction between a human user and intelligent conversational agent; furthermore, this section presents conversational agent development platforms that can be used train an affective conversational agent (e.g., DialogFlow).

Chapter 4 examines deep learning techniques and their application to natural language understanding tasks like affective conversational modeling. This chapter covers the selected algorithm and architecture in greater depth.

Chapter 5 examines the experimental evaluation of the current methodological approach (i.e., classification task methodology, generative conversational modeling task methodology). This chapter presents the quantitative and qualitative analyses of model performance.

Chapter 6 examines the implications of the current research aim and results in long-term research endeavors. Furthermore, the upcoming chapter will propose steps that can be taken to extend current research aims in upcoming phases of the investigation.

Chapter 2

Background and Related Works

Emotional expression through language is a highly complex phenomenon influenced by a variety of socio-cultural linguistic factors. For example, expressions of emotion are generally more indirect within collectivist cultural contexts than those within individualistic cultural contexts [7]. Furthermore, some emotions, like anger, are commonly expressed through the use of metaphors ("e.g., you make my blood boil"). Emotional linguistics is highly ambiguous. Thus, textual affect sensing continues to be a hard problem of artificial intelligence because the ambiguity of language, particularly emotional dialogue is a highly challenging computational task for a machine to tackle. A variety of textual affect sensing techniques have attempted to capture linguistic emotional complexities within recent years. These techniques have relied on previously developed representations of emotion. A few popular frameworks will be explored in the upcoming section.

2.1 Foundational Frameworks of Emotion

Currently, there are two dominant frameworks of emotion: discrete and dimensional representations. Robinson and Baltrusaitis [8] discovered that the traditional discrete representation of emotion can be problematic. This may be due to the fact that emotion, as experienced on a daily basis, is not just a discrete point in the (valence, arousal) space, see Figure 2.1.

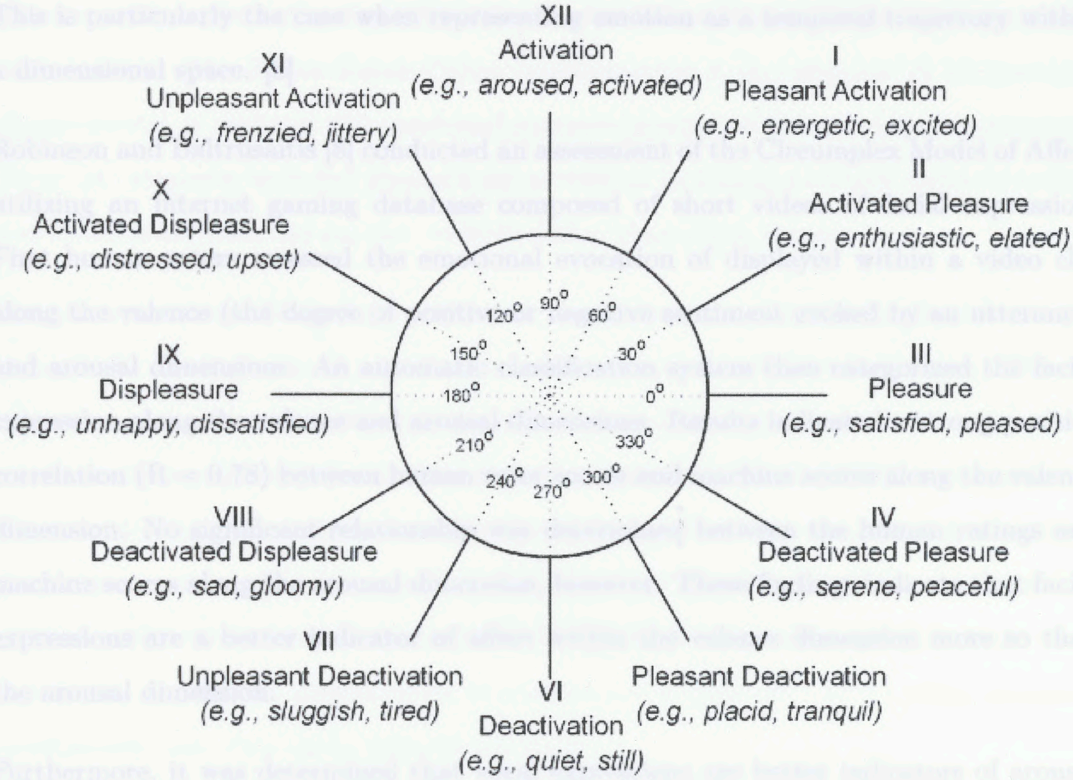


FIGURE 2.1: Russell's Circumplex Model of Affect; Notice here that the roman numeral annotations are labeling the twelve segments of Core Affect. For example, the segment between 30 degrees and 60 degrees falls within the Pleasant Activation (I) and Activated Pleasure (II) states.

They proposed an alternative way to represent emotion as a combination of two or more dimensions. The use of a continuous, dimensional emotional model was explored as in Russell's Circumplex Model of Affect, see Figure 2.1. Most of the previous research that has evaluated the use of such models have been limited to nonverbal expressions of emotion (i.e., facial expressions, audible utterances), so generalizability to the current research effort may be limited. Still, this research provides a powerful compass for determining the effectiveness of dimensional models as will be further explored in the following section.

2.1.1 Russell's Circumplex Model of Affect

Russell's Circumplex Model of Affect has been determined to be one of the most effective representations of emotion [4]. For temporal trajectories of emotion over dimensional space (e.g., videos), this model has been shown to especially effective representation.

This is particularly the case when representing emotion as a temporal trajectory within a dimensional space. [8]

Robinson and Baltrusaitis [8] conducted an assessment of the Circumplex Model of Affect utilizing an internet gaming database composed of short videos of facial expression. First human raters assessed the emotional evocation of displayed within a video clip along the valence (the degree of positive or negative sentiment evoked by an utterance) and arousal dimensions. An automatic classification system then categorized the facial expression along the valence and arousal dimensions. Results indicated a strong positive correlation (R = 0.78) between human rater scores and machine scores along the valence dimension. No significant relationship was determined between the human ratings and machine scores along the arousal dimension, however. These findings indicate that facial expressions are a better indicator of affect within the valence dimension more so than the arousal dimension.

Furthermore, it was determined that vocal expressions are better indicators of arousal than valence [8]. This research does not suggest how much textual affect sensing techniques will be influenced by the valence or arousal dimensions. However, based on findings from other studies (e.g., [9]), it can be inferred that valence is a more dominant determinant of overall affective quality of utterances (i.e., textual affect sensing tasks). Additionally, Robinson and Baltrusaitis [8] suggested that "universal emotion meters" are generally inadequate. They suggest that a domain specific model of emotion should be used within a specialized contextual application, given the large range of dimensional overlap between specific emotions. Thus the current research aim will be to create Army-based professional conflicts dialogue corpus that can be used to support emotion classification and dialogue generation tasks. Furthermore, the methodological approach will be scenario/vignette specific, to assure high domain specificity.

Previous models of emotions, such as the one proposed by Charles Darwin in *The Expression of Emotions in Man and Animals* [10], later revised by Paul Ekman represent emotions within 6 discrete categories (joy, anger, disgust, fear, sadness, and surprise). Ekman's model has received particular popularity for affective computing tasks over the past four decades. However, these emotional states are not reflective of daily emotional experiences, given that more subtle and ephemeral mental states are experienced within daily interactions. Broader taxonomies of emotion have been introduced in an attempt to acknowledge this limitation. This includes the Simon Baron-Cohen linguistic analysis taxonomy. The Simon Baron-Cohen representation is an extension of the previous Ekman model, it contains 412 emotional concepts grouped within 24 disjoint categories. These 24 categories included Ekman's six as well as 18 complex mental states identified over continuous observations (i.e., videos) rather than static images.

Furthermore, James Russell [4] addressed the limitations of previous discrete models through the development of a continuous, dimensional classification approach. The developed model was created through the task of placing 28 emotion-derived words around a unit circle representation. A Principal Component Analysis (PCA) was applied in order to identify major dimensions within the data. The PCA uncovered two primary dimensions, valence (located upon the horizontal axis) and arousal (located upon the vertical axis). Robinson and Baltrusaitis [8] concluded that this model is more effective for the continuous identification of emotions on a scale of -1 to +1 when measured continuously over time using video frame rates.

While the findings of this study demonstrate the effectiveness of the Russell Circumplex Model representation of emotion there are limitations to be considered. First, the analysis involved the observation of facial recognition video datasets. The currently developed model will be applied to a textual affect sensing task. Assumptions about the effectiveness of the model in application to the currently proposed task cannot be made. Furthermore, the low correlation values for assessment arousal scores indicate that the use of the Circumplex model may render to the inadequate identification of the intensity or arousal scores for the currently proposed framework.

On the other hand, this study demonstrates the effectiveness of Russell's model for the recognition of emotions over temporal trajectories. This will be particularly pertinent to the current research given the usage of the sequence to sequence architecture. The sequence to sequence architecture represents the conversational temporal trajectory using implicit vectorization of the given conversation. Overall, results indicate that the Circumplex Model of Affect will be an effective framework to apply to the given computational task especially within the valence dimension.

Posner, Russell, and Peterson [11] conducted a meta-analytical exploration of the application of the Circumplex Model of Affect to existing behavioral, cognitive science, neuroimaging and development theories. Researchers defined emotion as a neurophysiological linear combination of two dimensions of varying degrees valence and arousal. Emotions arise from activation within unique neural pathways (i.e., mesolimbic, reticular). Emotions are hence a continuum of overlapping states. As suggested by previous research there is often very little differentiation within the arousal dimension, especially for emotions like (anxiety and depression). Furthermore, findings indicated that emotional states are not always characterized by affective behavior changes (e.g., anxious state not always demonstrated by changes in facial expression). These findings illustrate how some emotions may not be explicitly identifiable using the currently developed textual model.

2.1.2 Overview of the Circumplex Model of Affect

Willigen [12] represented emotional states using the Circumplex Model of Affect shown above. Figure 2.1 illustrates the representation of emotional states using the Circumplex Model of Affect [13]. This model conceptualizes affective states as the product of two neurophysiological systems, valence (levels of pleasantness or unpleasantness) and arousal (levels alertness or activation). Each affective state is interpreted as a combination of these two spectrums. For example, elation may be a combination of a high positive valence and a moderate arousal level within the two aforementioned neurophysiological systems [11].

2.2 Frameworks of Emotional Transitions

Thorton and Tamir [14] examined the accuracy of emotional transition predictions using mental models. Findings suggested that human beings are highly adept to the predictions of emotional transitions up to two transitions into the future [14]. Raters predicted many transitions with high accuracy (e.g., transitions from feeling touch to distress [and vice versa], transitions from gloomy to sad states, transitions between emotions with similar valence (e.g., sad and gloomy) were predicted as most likely to occur). Although four main conceptual dimensions were identified as pivotal determinants of emotional transitions (i.e., valence, social impact, rationality, human mind [overall holistic similarity]), the process was primarily indicated as being heavily intuitive and reliant on inherent mental models.

Results unveiled an illuminating phenomenon, emotions predict emotions (i.e., current emotions predict future emotions) and furthermore emotions generally predict actions (e.g., tired people rest). The ability to predict emotions is a highly beneficial skill to possess as social creatures (e.g., can predict the likelihood of career success, family interactions, skill acquisition). Hence, the best source of models for emotional dynamics is pointed towards human mental models. Thorton and Tamir [14] utilized Markov modeling to explore temporal models of emotion. First, an estimation of real-world mental states was identified. The accuracy of the mental models was assessed using correlational and normal root mean square error (NRMSE) analysis. This analysis revealed strong associations between average mental models and experience samplings (e.g. Spearman R of 0.77) over 60 different emotions. This study demonstrates the high accuracy of human mental models of emotion transition.

Overall these findings indicate that human beings are the primary experts for emotional transitions. While these findings are highly illuminating to the currently proposed research questions there are a couple of limitations. First, the analyzed data was mainly self-report data. The currently developed model proposes the extraction of emotional dynamic trajectories using Natural Language Processing (NLP) vectors. Secondly, we can not determine how applicable the findings are to emotional transition within a text-based chatbot model.

Furthermore, these results are particularly informative to the currently developed model because they demonstrate that emotional transition predictions should be dyadic and that transitions should take place within the same valence range). Hence, the arousal dimensions may be as informative to the process of determining emotional transitions (through implicit sequence analysis) vectorization as the valence dimension.

The current study will utilize these findings through the development of a Real World Professional Conflicts Dialog Corpus. This corpus will be compiled through a process of collecting predictions of emotional transitions, in response to various professional conflicts, from human participants.

2.2.1 Emotion Dynamics

Emotional inertia is the degree to which an emotional experience transfers over from moment to moment, while inter-speaker emotional influence is the degree to which one person's emotions influence another individual to state [15]. Both of these concepts play a vital role in the nature of emotional transitions and thus should be of consideration when informing the development of the current model. It is important to distinguish these two factors of emotional dynamics given that emotional experiences are not solely subjective but are subject to change based upon socio-environmental factors (i.e., given mental state of dyadic conversational counterpart). The currently developed framework will assume that emotional dynamics are mirrored between dyadic counterparts (e.g, speakers tend to match the emotional states of their speaking partners) [14].

2.3 Traditional Approaches to Textual Affect Sensing

One of the most popular traditional textual affect sensing techniques is keyword spotting. This technique categorizes utterances into affective classifications using words with obvious emotional salience (e.g., "happy", "despondent", "frustrated") [16]. While this technique is popular and relatively easy to carry out, results may be poor and limited 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 in actuality. 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 [16].

Notably earlier applications focused on the development of relational agents. For example, Bickmore and Gruber [17] explored the development of a relational agent that was used for "health counseling and the encouragement of health behavior changes (e.g., medication compliance)." The system's ability to build rapport with patients was also evaluated. The developed relational agent was able to evaluate patient compliance levels and thus alter the communication style in order to foster patient understanding of the necessary interventions. However, this system utilizes multiple choice selection instead of free response input. The use of such a "forced-choice" conversational system could introduce many of the limitations presented by traditional assessments [3] if implemented within the currently developed framework.

Lui, Lieberman, and Selker [18] explored a novel approach to textual affect sensing using a real-world common sense knowledge base. An open mind common sense 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 (composed of six primary emotions [i.e., happy, sad, angry, fearful, disgusted, surprised]). The approach went beyond the traditional surface-level feature extraction techniques seen in traditional approaches to textual affect sensing (e.g. keyword spotting, Statistical Natural Language Processing, lexical affinity, rule-based methods), analyzing underlying semantic meanings related to affect on a sentence level. This developed system was embedded in an email browser. Users reported that the implementation was robust enough to be used in day to day communication (i.e. emails). Metric-based outcomes, however, were ambiguously presented. Researchers utilized generic common sense knowledge database containing English sentences like the following: "Some people find ghosts to be scary"; "A consequence of riding a roller-coaster may be excited. [18]" These sentences were analyzed using linguistic processing that facilitated subject-verb object identification as well as semantic processing. The use of emotion ground keywords allowed for emotional classification upon the six Ekman emotion model. Finally, an affective valence score was computed utilizing a propagation trainer. The text analyzer was composed of five main modules: text segmented, linguistic processing suite, story interpreter, smoother, an expresser. Disambiguation metrics were used to generate a final emotion annotation score after utterances passed through these five main components.[18]

A major limitation of this study is the modularity of the system architecture. In the case of this system architecture, affect sensing operates independently of the user and story contexts [18]. Furthermore, while this approach allows for advances in affect sensing, affect understanding remains unexplored. The currently developed system will address this weakness through the use of end-to-end encoder-decoder system architecture (i.e., sequence to sequence). This may lead to a greater performance on textual affect analysis and textual affect generation tasks. Furthermore, through the use of a popular conversational corpus, Cornell Movie Dialogue Conversation Corpus [19], the currently developed system will be allowed to use the strength of a common sense knowledge base in an unconventional and implicit manner.

Although there are a plethora datasets for sentiment analysis (e.g. [20]), these datasets may often lack generalizability to emotion-based tasks. Mohammad and Turney [21] have addressed this problem through the use of crowd-sourcing platforms like Amazon Mechanical Turk. They developed an emotionally annotated dataset of 14,000 English words. An existing highly popular lexicon, WordNet-Affect, has been used in many applications of sentiment analysis, opinion mining, and emotion detection. The Depeche Mood emotional lexicon [22], LIWC [23], and ANEW (Affective Norms for English Words) [24] are three other lexicon based datasets that have been annotated based on dimensional models of emotions, dimensions including valence, arousal, and dominance. Vector Space Modeling has offered some ways of increasing performance of emotional classification models (e.g., [25]). The methodology was based on the deconstruction of semantic models through word embeddings. Within this model, the individual words are represented as vectors of n-dimensional space. Within this vector, the distance between vectors corresponds to the level of semantic similarity between words. This modeling technique has been applied to machine translation, Named Entity Recognition (NER) and has produced modest to robust results.

Supervised Learning approaches to textual affect sensing have faced a variety of challenges, mainly related to the lack of high-quality balanced data-sets. Some research has mitigated this limitation through the use of social media posts and microblogs, most commonly Twitter data (e.g., [26], [27]). Emoticons, hashtags, and emojis have all been used as labels that support supervised training (e.g., [28], [26]). Previous approaches have utilized three main frameworks of emotion to support classification including Plutchik's Wheel [29], Ekman 6 Emotional model [18], and Russell's Circumplex Model of Affect [30]. A variety of training methods have been explored. Many of these techniques are within the traditional Bag of Words (BOW) domain (e.g., Support Vector Machines [31], Support Vector Regression [e.g., [32]). Others have mainly explored lexicon approaches (e.g., LIWC (Linguistic Inquiry and Word Count), MPQA, WordNet-Affect, POS) [7]. KNN and Decisions Trees have also been applied to the analysis of emotion within the text. Such approaches have produced mediocre to modest results.

Suttle and Ide [33] created a binary classifier using the Plutchik's wheel and used a manually annotated twitter dataset with hashtags, emoticon, and emojis as labels. Results varied between 75 percent and 91 percent accuracy. This discrepancy in performance could mainly be attributed to the differences in the amounts of high-quality data for various emotions. For example, tweets that depicted anger and happiness were more common than those that depicted surprise. Still, these results are highly promising and demonstrate the benefits of using microblog data to support training. Other approaches have produced results with high levels of disparity between under-represented and highly representative emotions. Purver and Battersby (2012) developed a Support Vector Machine (SVM) classifier. Results were moderately strong for happy emotions (82 percent) but were as low as 13 percent for other underrepresented emotions. Balabantaray [34] addressed this disparity by identifying a plethora of features to extract from manually labeled data (e.g., Unigrams, Bigrams, Personal-pronouns, Dependency- Parsing).

Particularly pertinent to the current research efforts are results indicating the robustness of unigram models for the identification of emotions within the Circumplex model framework (i.e., within the four main emotional categories). Results have produced accuracy rates close to 90 percent (i.e., [35]) and have exploited a variety of techniques in combination with Naive Bayes, Support Vector Machines, Decision Trees, and KNNs.

Furthermore, some of the previous approaches to textual affect sensing have included unsupervised learning techniques. Many of these techniques have utilized dimensionality reduction methods (e.g. Latent Semantic Analysis [LSA], Probabilistic Latent Semantic Analysis [PLSA], Non-Negative Matrix Factorization [NMF]). Instead of the traditional emotional frameworks(e.g, Circumplex Model of Affect, Ekman, Plutchik's Wheel), unsupervised learning approaches have utilized other dimensional models (e.g., ANEW [Affective Norms for English Words], WordNet-Affect, NAVA (Nouns, Adjectives, Verbs, Adverbs). For such approaches, emotional assignments can be based upon the proximity or level of syntactic dependency between words within the vectors (cosine similarity, Point-Wise Mutual Information [PMI] Measures). These approaches have generally produced modest performance outcomes (e.g., [36].

Rule-based NLP systems have demonstrated robust performance within the realm of unsupervised learning. Rules that distinguish linguistic language patterns and the Rule-based Emission Model have been used in combination with LIWC (Linguistic Inquiry and Word Count) Lexicon and have produced results comparable to the state of the art supervised learning vector-space modeling techniques (e.g., [37]).

2.4 Novel Approaches to Textual Affect Sensing

2.4.1 Examining a Traditional Deep Learning Approach

The Microsoft research lab explored the application of a simple deep learning algorithm to the textual emotion detection and recognition task and yielded modest results. Researchers addressed the lack of high-quality datasets through the use of human-expert labeling services like Mechanical Turk [38]. Mechanical Turk (MT) is a low-cost service that recruits workers to perform tasks that require some level of human intelligence (i.e., in this case labeling the emotional categorization of an English phrase or utterance) [38]. For this study, MT raters determined the regressional levels of each emotion presented in the text upon a continuum.

Furthermore, the model was trained using a dataset that consisted of 784,349 samples of informal short English messages (tweets) within five emotional categories including the following: anger, sadness, fear, happiness, and excitement. The training to validation to testing ratio was 60 percent: 20 percent: 20 percent, respectively [38]. This dataset was formulated using several sources of data including the ISEAR database and the SemEval 2007 database, two of the few, existing emotional text databases with labeled data. The remaining data was extracted from product reviews, journals, fiction excerpts, and news articles and then labeled utilizing Mechanical Turk services [38].

While the overall results were not state of the art, the unweighted accuracy of emotional recognition through text was 64.47 percent. Findings indicated that fear was the most difficult emotion for the model to classify accurately. On the other hand, the model did moderately well in recognizing emotions like anger, sadness, and excitement. The developed artificial neural network model consisted of three layers. The first layer has 125 neurons, the second, 25 neurons, and the third, five neurons [38].

There are several limitations presented by the results of this study. [38] The sample of MT raters may have consisted of a small number of individuals with similar linguistic and cultural backgrounds. The emotional rating within the datasets may lack external validity, for this reason. Perhaps a more high-quality dataset could have led to higher accuracy rates. Another possible limitation was the selected training model. While the use of a general deep learning architecture yields modest results, some studies (e.g., cite Seq2seq here) have shown the use of an end-to-end Recurrent Neural Network(RNN) based algorithm generates more high-quality results more consistently (e.g. classification accuracy rates of 85 percent or greater). This study was conducted in 2015 and since that time more and more research has surfaced that has produced more sources of high quality annotated emotional databases of text (e.g., Sem-Eval 2018, Clean-Balanced Emotional Tweets [CBET] dataset). Current and future work may utilize similar sources of emotionally rich annotated data [38].

2.5 Sequence to Sequence Architecture

Sutskever, Vinyals, and Le [1] examined the application of a novel model to NLP deep learning tasks. They discovered that the performance of existing translation models using recurrent neural networks could be augmented through the use of the Sequence to Sequence Architecture (seq2seq) [1]. A state of the art performance on an English to French translation task was achieved. The overall performance score was within-in five points of the existing state of the art system ([39], achieving a BLEU score of 34.81). One of the identified limitations of this study's findings was the inability of the model to handle words that were not presented in the training sample. However, the strengths of this study greatly outnumbered potential presented weaknesses.[1] Unlike previous models (e.g., Bag of Words Model) this model was semantically sensitive to word order and was able to represent new meanings for vectors with the same elements and varying degrees of order within the presented vector. Additionally, the proposed system was able to robustly handle both relatively long utterances and short utterances [1].

The Long-Short Term Model (LSTM), as well as the use of attention mechanisms, support long-range conversational flow [1]. It is important to note that this approach was not specifically oriented toward the contextuality of emotion through language but rather focused on the improvement of a generalized contextual conversational agent. Still, the findings of this study are relevant and applicable to the current research project.[1] The currently developed framework will need to utilize a system that is able to handle the contextuality of emotional utterances throughout a conversation in order to preserve meaning beyond a specific conversational domain.

2.6 Sequence to Sequence Approaches to Textual Affect Sensing

Mohammad, Marquez, Salameh, and Kiritchenko (2018) [26] examined a previously developed approach to Affect Sensing in Tweets ([40]). A multilingual dataset was utilized (i.e., datasets for English, Arabic, and Spanish). The dataset was assembled using a variety of tools including DeepMoji [41], a neural network used for matching emojis to expressions in tweets. Altogether a dataset of over 22,000 multi-labeled tweets was assembled and used for training 5 comparative machine learning models on emotion analysis tasks. Tweets were analyzed for both coarse-level and fine-grained level of effective content. Finely grained scores were generated using Best-Worse Scaling (BWS) and were shown to have high levels of validity (e.g., split-half reliability greater than 0.80). A comparative examination of two machine learning approaches was conducted. It was determined that overall, deep learning approaches outperformed the traditional SVM-unigram approaches. The top-performing models related to deep neural network representations but performed particularly well when combined with manually engineered features, common among traditional approaches (e.g, features derived from affect lexicons) [41].

While deep learning representation learning approaches were indicated to be highly effective, it was determined that lexicons like ANEW(Affective Norms for English Words) ([24]) and the NRC Emotion Lexicon ([42]) greatly improve the performance of generative deep learning models.

Overall the system's methodology yielded robust results. Particularly pertinent to the current research endeavor was the use of a regressional-dimensional emotional modeling framework. In fact, researchers conducted emotional analysis tasks in 5 main separate steps instead of one conglomerate step (i.e., 1. Regression analysis of emotional intensity; 2. Classification of emotional intensity 3; Regressional analysis of valence; 4. Classification of emotional valence; 5. Classification of emotion). Given that the current study will focus on the use of a deep learning representation upon a dimensional framework of emotion, these results are particularly promising. However, the current study will not utilize a deep learning model in conjunction with a traditional machine learning algorithm (e.g., Support Vector Machine/Support Vector Regression approach, as seen in lexicons like PlusEmo2Vec). Thus performance outcomes may have limited applicability to the current study. Furthermore, it was not particularly clear whether the models with Recurrent Neural Network (RNN) layers or Convolutional Neural Networks (CNN) layers yielded more robust results [26].

Honghao, Zhao, and Ke [43] developed a generative chatbot model using Seq2Seq architecture. Unlike typical Seq2Seq approaches, however, the system architecture was augmented with internal and external memory modularity. Overall, this system was determined to generate "reasonable responses." This study utilized a categorical approach to emotion analysis, that was based upon the Ekman 6 emotion framework. Eighty-eight million subtitles of movies and TV programs were retrieved for model training. Another distinguishing characteristic of this study was the use of a pre-trained model for word vectorization along with the co-training word vectorization typical of seq2seq models. The quantitative outcomes of this study were unclear. However, it is clear that results were promising and indicated the strength of sequence to sequence architecture when applied to the generation of emotionally contextual responses [43].

Zhou et al., [44] extended the previous investigation of the generation of effective dialogue through the proposal of novel architecture, namely the Emotion Chatting Machine. This architecture consisted of three main mechanisms, including the following: Emotion Category Embedding Mechanism, Internal Memory Mechanism, External Memory Mechanism. In contrast to some traditional approaches this approaches, that focus primarily on emotion detection and classification, this approach focused on providing the generation of contextual responses (within an emotional domain). Overall system evaluation revealed that the model successfully accomplished the task of generating emotionally contextual responses. This system achieved a perplexity score of 65.9 compared to the conventional seq2seq architecture which achieved a perplexity score of 68.0. Overall this system attained an accuracy rate of 77.3 percent compared to the accuracy produced through the use of the traditional(vanilla) seq2seq model, 17.9 percent. Particularly the use of a knowledge corpus for the course-grained detection of emotion in conjunction with an internal memory module may have augmented the performance of the proposed model [44].

Perhaps the currently proposed system should seek to implement a memory module within current and future versions of the proposed model [44]. Although the Emotion Chatting Machine architecture is highly state-of-the-art, it should be noted that the findings of this research may have limited generalizability to the currently proposed model given that the utilized dataset was composed of mainly Chinese (Weiblo) blog posts. Furthermore, a discrete categorical emotional representation (Ekman's Framework) was utilized. It should be considered that current, efforts will utilize dimensional emotion models. Particularly pertinent to the current study is the high level of focus on the generative module of the system architecture, and thus the system's overall ability to produce emotionally appropriate responses [44]. The current model will contribute to research efforts that ultimately seek to accomplish a generative dialogue task, not just emotion detection and analysis in isolation. Thus, the developed system produced a variety of responses corresponding to the intended emotion for an inputted utterances like, "Best day ever. I just had a huge piece of my favorite chocolate fudge cake!"

Gee and Wang [45] expanded a previous implementation of emotional analysis tasks of tweets [26], through the use of transfer learning. The WASS-2017 Shared task of emotional intensity dataset was utilized in this research effort [46]. They found that the transfer learning from sentiment tasks allowed the system training to overcome the general lack of emotionally labeled training data. The utilized approach, transferal of knowledge from sentiment to emotion, is relatively unique and novel but demonstrates the promising potential of utilizing such an approach in the development and testing of the current model.

Gee and Wang [45] utilized a previously trained model (i.e., [26]) and combined the penultimate layers into a single vector, through the use of hierarchical clustering. Multi-dimensional word embedding in conjunction with a single dimension lexicon base features allowed for the improvement of the performance of previous implementations. The developed system outperformed traditional systems that train specific emotions independently.

Zahiri and Choi [47] introduced a new corpus for emotion detection utilizing spoken dialog from the television show, *Friends*. They proposed an attentive Sequential Convolutional Neural Network (SCNN) model. It was revealed that this approach outperformed the base Convolutional Neural Network (CNN) on emotion detection tasks. Researchers chose to use a SCNN model due to the inability of traditional CNN approaches to take the historical context of utterances into account. While RNNs are traditionally highly adept at historical contextual tasks, this methodology was utilized due to the fact that RNNs generally require a hefty amount of data in order to prevent model overfitting as well as the rather slower training performance of RNNs. Through the use of the SCNN approach, the system attained an emotional detection accuracy of 54 percent. This accuracy is higher than the CNN baseline but is still less robust than some recent RNN approaches. These approaches will be further explored in the next section.

Overall, RNN methodologies have achieved more robust performance outcomes on textual affect sensing tasks. Mageed and Ungar [48] achieved a state of the art performance on a system that analyzed 24 fine-grained categories of emotion, and produced an average accuracy rate of 87.58 percent. They extended those findings through the application of Robert Plutchik's Eight Primary Emotion framework and achieved a superior overall accuracy of 95.68 percent (coarse-grained emotion analysis).

This research effort augmented model performance by addressing the absence of large labeled datasets and demonstrating the superiority of Gated Recurrent Neural Networks (GRNNs), in particular, on emotion detection tasks. One potential limitation of this study is that it utilized a different framework than the selected dimensional emotion framework utilized for current research endeavors (Circumplex Model of Affect). Perhaps, for this reason, findings may not be as applicable to the current research aim. However, this is unlikely given the use of large high-quality dataset. Furthermore, it should be noted that the GRNN architecture may perform significantly better than RNN approaches if applied to methodologies of current and future development efforts of the proposed framework [48].

Shirai et al. [49], conducted a comparative analysis of several deep learning algorithms on textual affect sensing. This analysis was conducted through the development of complaint classification systems. Several RNN models were utilized including the following: FNN, GRU, LSTM, GRU-GRU, LSTM-GRU, GRU-LSTM, LSTM.

The FNN model outperformed most of the others by a slight margin (i.e., $p = 0.859$, $r = 0.856$) but was followed by the GRU (i.e., $p = 0.847$, $r = 0.845$) and LSTM (i.e., $p = 0.857$, $p = 0.55$) approaches. It should be noted that findings from this study may have limited generalizability due to the use of a non-English dataset (consisting of utterances in Thai) [49].

Majumder et al. [50] analyzed dyadic emotion conversations recorded through video. They utilized the conversational memory network architecture to conduct this analysis. This model augmented previous state of the art approaches (e.g., LSTM) which generally perform poorly on long-range summarization tasks, though they have robust overall performance (e.g, [51]). They developed a system capable of handling long-range emotional trajectories through the use of contextual memory networks, that is a continuous vector that keeps a historical recollection of contextual cues within memory cells. The use of an attention module further augmented the performance of the model. Overall this model outperformed the existing state of the art approach (increased accuracy rate by three to four percent).

While the findings are informative to the current study, the analyzed dataset consisted of video-based conversations. The current model will examine datasets that contain textual affective utterances, thus these results have limited applicability.

2.7 Limitations of Previous Approaches to Textual Affect Sensing

Textual Affect Sensing continues to be a hard problem of Artificial Intelligence [1]. Emotional representations as a framework continue to be a difficult task within various fields of psychology. Previous attempts to determine emotional salience through text have focused heavily on rule-based and Support Vector Machine (SVM) approaches. While these approaches have yielded modest results, the outcomes have been limited, due to the lack of consideration for the order of presented words in a vector. The currently proposed model seeks to extend previous research in the area of textual affect sensing, by applying a recently developed approach to natural language processing, deep-learning focused sequence analysis. The current approach will address these limitations by leveraging the strengths of the seq2seq architecture [1].

2.8 Overview of Currently Proposed System Architecture

This architecture is the current state of the art for natural language processing tasks and has been used in recent years for the improvement of textual affect analysis and conversational models. There are several pertinent benefits to applying the seq2seq architecture to the current system. First, unlike the previous machine learning architectures, seq2Seq allows for robust performance outcomes on both relatively long utterances (e.g., greater than or equal to 255 characters) and relatively short utterances (e.g., shorter than 255 characters). This will be a particular benefit for the current study seeing that users of the interactive educational system under long-term development may offer utterances of varied length. While it is preferred that the utterances should have a brief conversational length. It is beneficial to have system architecture should be prepared to handle a variety of verboseness among inputted utterances.

The most pertinent benefit of the proposed system architecture is that the seq2seq architecture would allow the model to vectorize inputted utterances in a way that takes linguistic ordering into account. The previous Bag of Words models have accounted for the frequency of words in the "system vocabulary" more so, than the contextual cues between and surrounding those words. The current system will seek to address the limitation by utilizing a Recurrent Neural Network approach. Recurrent Neural Networks are Artificial Neural Networks that feed the output back into the network recursively. The use of such networks allows for the contextual remembrance between neuronal network layers and thus the analysis of input in a sequence (e.g., the sequence of words in the text). This will be a particularly important strength given that future uses of the developed model at the Army Research Institute will enhance currently developed interactive learning simulation software, which presents scenarios to junior officers. Junior officers will be required to provide verbal utterances that demonstrate empathetic leadership styles to virtual agents within interactive scenarios. It is expected, that in the future the virtual agents will remember emotional context between scenes (e.g., a virtual character will keep a historical recollection of the previous user utterances and the corresponding influence on the agent's state within a previous scene; the agent may offer an utterance that will reflect the resulting emotional state). The use of the Seq2Seq architecture will allow for the representation of an "emotional memory" of virtual characters in future research endeavors.

2.9 Overview

Overall, traditional textual affect sensing expert methodologies have produced mediocre to modest results. This can be attributed to several limitations. Linguistic emotional expression is a highly complicated phenomenon. The use of metaphorical expressions and implicit context dependencies make linguistic expressions of emotions even more ambiguous. This is still a hard-task for machine learning algorithms and has yet to be fully explored in current and past bodies of research. Secondly, high-quality datasets corresponding to contiguous emotional models (e.g., Circumplex Model of Affect). While micro-blog datasets (tweets) have allowed for the mitigation of limitations related to the expressions of emotions on social media are not necessarily characteristic of day to day emotional expression seen in everyday dialogues. There is still a high need for annotated conversational data with balanced amounts of data for various emotions. Lastly, current models of emotions are inefficient, most of them are limited to Bag-of-Words representations. Such limitations have relied more heavily on the content of words within a vector rather than the relationships between words in those vectors. This is a huge limitation given that linguistic emotional expression is highly contextual.

More novel approaches may utilize neural networks in order to fully characterize those dependencies within emotional analysis tasks. LSTM-based Recurrent Neural Networks, which allow for the "historical recollection" of contextual cues throughout the conversational vectors, can be particularly useful in the accomplishment of the Natural Language Understanding (NLU) based textual affect sensing tasks. The goal of NLU is to uncover deeper semantic meanings within text [6]. Chapter three will overview recent deep learning approaches to conversational agent development.

Chapter 3

Conversational Agents

Dialog has become a common interactional medium between human users and machines in recent years [52]. Some examples of popular or commonly known chatbots include personal home assistants Amazon Alexa, Amazon Echo, and Google Home. A few of the advantages offered through conversational agents include the decrease in cost for customer service resources among major companies, the use of an intuitive interface with which more users may be more familiar (e.g., 'Alexa open YouTube,' instead of fire-stick buttons). Furthermore, due to recent advances (e.g., deep learning approaches to sentiment analysis (e.g., [26]), recent approaches to information delivery (e.g.,[52]) in conversational agent development, delivery has become more personalized (e.g., 'Traffic is low this morning, it will take you 11 minutes to arrive'). Due to these advances, conversational agents are getting closer to passing the infamous Turing Test, simulating a human conversation convincingly and realistically. Some may argue that this feat has already been accomplished. The current research aim is to model the simulation of empathy-based human to human conversation [53].

Conversational agents are dialog systems that mimic realistic interactions between people [53] [54]. They may embody conversational agents (e.g., be presented as the avatar, humanoid), be traditional voice-based (systems that receive mainly sensory speech input), or text-based systems (receive mainly text input). One type of conversational agent is the chatbot. Chatbots receive natural language input and utilize this to produce a goal-directed output on behalf of the human user. Chatbots are considered to be both social agents and intelligent agents [53] [54].

Intelligent agents are autonomous, reactive, social and most importantly adaptable [54] [53]. Chatbots that are adaptable are intelligent and are to learn from previous experiences (i.e., dialog exchanges). Generally, intelligent conversational agents can generate novel responses given novel inputs. They also train them to respond differently to those utterances in the future. Machine learning algorithms (e.g., Markov, Deep Neural Networks) have allowed for the emergence of state of the art chatbots that anticipate future interactions from previous interactions with human users. However, chatbots have not always been adaptable [54][53].

Early approaches to chatbot development focused on the use of hand-crafted rules, keyword matching, and minimal context identification techniques. [53] These systems (e.g., ELIZA, ALICE (particularly popular for a pattern matching XML based chatbot development language, and Artificial Intelligence Markup Language (AIML)) were exemplary stepping stones within the field of conversational agent development because they attempt to pass the Turing Test. However, when conversing with many of these earlier systems, it was clear to most human users that conversation was taking place with a machine instead of a person. While today's chatbots have not been unanimous "Turing Test approved," great advances have been made that provide more realistic and fluid interactions between humans and machines. For example, this is used to deceive users on online social media platforms, like Twitter (e.g., Sybil bots) to propagate malicious software [53].

The interaction between chatbots and human users are determined by the use of the conversational interface [54]. There are two main types of conversational interfaces: transactional and conversational. Transactional chatbots are utilized for the accomplishment of specific tasks (e.g., "place an order for matcha green tea on Amazon," "show me the movie schedule for today"). Conversational chatbots are mainly used for social ("chit-chat") purposes. Chatbots may be either text-based or voice-based. Overall, these interaction styles determine the delivery of human-centered services and responses. To support this style of interaction, chatbots must be capable of more than rudimentary processing of natural language, going a step further to extract meaningful information from presented utterances. This is why Natural Language Understanding continues to be a related task in the development of conversational agents [54].

3.1 Natural Language Understanding for Conversational Interfaces

Many modern natural language toolkits are utilized for Natural Language Understanding (NLU) tasks [54]. These toolkits extract intents and entities for natural language input. The intents are actions or goals that the user seeks to accomplish (e.g., "Request an Uber ride," "Order my favorite three topping mushroom pizza"); entities are parameters necessary to achieve those tasks.

Intent recognition is typically a machine learning task (e.g., traditionally Support Vector Machine[SVM] / Bag of Words (BoW) Classification) [54]. For these machine learning tasks sample, utterances are introduced to the machine, and the similarities and variations among those samples are determined.

As previously mentioned, intents are parameters that allow leverage intent recognition tasks. Examples of intents include location, time, and pizza toppings. Recurrent Neural Networks have been particularly useful for entity extraction in recent research (e.g., [55]).

3.1.1 Conversational Interaction Styles

The requirements of NLU tasks often guide conversational interaction. There are four main types of interactional styles between users and agents: one-shot queries dialogue, slot filling dialogue, mixed interaction, and open-ended dialogue)[54]. One shot queries are user-initiated and occur in the simple input-output pair format (e.g., "play a song that will lift my mood"). Slot filling dialogue systems collect information to fulfill requirements for user responses(e.g., "User: schedule an Uber ride," "System: What time?", "User: 9 a.m.").

For this interactional style, follow up questions are useful because they leverage the collection of pertinent information for entity extraction. Mixed dialogue systems are both system initiated (as are slot filling dialogue systems) and user-initiated (as are one-shot queries). Developmental efforts may be centered upon the right balance between the two [54].

Open-ended dialogue systems strive for a blend of interactional styles (e.g., mixed initiate, multi-turn, multi-contextuality maintenance) and allow for more fluid, conversational flow. These are the most popular systems in current use today (e.g., Google Home, Siri, Amazon Alexa). There are currently popular platforms for open-ended dialogue system development (e.g., Dialog Flow, Motion.AI, Amazon Alexa Developer, Microsoft Bot Framework/Cortana, and IBM Watson Conversation). While they are all useful and adept at NLU tasks, DialogFlow is the platform utilized within current research efforts. Furthermore, this will be allowed due to the low cost to users and the natural developmental interface style [54].

3.2 Using Google's DialogFlow Dialogue Platform

DialogFlow is a Google-owned chatbot development platform that facilitates the augmentations of the human-computer interaction through conversation [54]. This company was founded in 2010 and bought by Google in 2016. Dialog Flow is used to create conversational interfaces for an eclectic range of purposes (e.g., phone apps, wearables, smart devices, and customer service chatbots). The robust support of voice-recognition, text to speech tasks, and text-based text speech conversion tasks make this platform an ideal choice for the simplified development of specific task chatbots.

Dialog Flow has a plethora of strengths such as easy and natural invocations (e.g., can invoke chatbot to begin a conversation the same way you would a friend, the flexible use of intents (i.e., context, user utterances events, action, and responses); robust training of intents (i.e., bot can recognize specific intents from a wide range of presented utterances with very little training data). The flexible use of entities (e.g., parameters identified by the agent during a conversation [i.e., time, location, weather]); increased flexibility of agent responses provided through the use of fulfillment requests [54].

Along with these benefits, the system has an intuitive setup for users with an existing Google account. In addition, it has a free non-enterprise edition available to all customers [54].

DialogFlow offers a cheap, quick, and effective chatbot development environment [54]. DialogFlow does all the heavy lifting on the front-end side of development to facilitate testing and simulation on the back-end development side.

Most importantly Dialog Flow, being Google-owned, offers similar benefits as other Google natural language processing platforms (e.g., DeepMind) (robust and powerful state of the art mechanisms for extracting semantic meaning from language) [54]. Intent recognition is still powerful and highly accurate with relatively low sample size. After intent training, the system can recognize valid intent, handle errors, and recognize similar natural language utterances while also retrieving valid responses for the user. Furthermore, the agents contains pre-defined default intents (e.g., parent intent, follow up intent) as well as fallbacks for queries (e.g., yes/no, cancel) [1] [54].

Dialog Flow allows simultaneous deployment of multiple specialized domain chatbots to conduct analytical comparisons of the datasets.[54] DialogFlow will support the current research in the following ways: all the determination of the proportion of validation data to train data in the production of optimal accuracy rates; allow for observation of alternative conversational modeling techniques for future usage in future tools [54].

This model provides a human-level conversation with the agent (i.e., Tensorflow-based agent predicted to be more likely to pass the Turing test). Unlike the traditional NLP model, this particular model will be a generative natural language model. Through the use of this model; it will be able to generate novel responses to novel inputted-utterances [54].

3.2.1 Other Approaches to Conversational Agent Development

As previously mentioned in Chapter 2, previous approaches to conversational agent development have been costly, error-prone, and lack generalizability to an external domain (e.g., handcrafted rule based chatbots like ELIZA). Statistical conversational agents leverage unsupervised reinforcement learning and Markov Decision Processing (POMDPs); however, they suffer weaknesses of having intractable decision sample spaces. End to End deep learning approaches are the most recent developmental efforts. They produce large corpora of dialog word by word and require large amounts of data (typical among supervised machine learning training tasks). The current research effort examines the use of this type of end to end conversational modeling, that is the application of the sequence to sequence architecture (seq2seq) [54].

3.2.2 Quality Attributes of Chatbots

Several factors determine the quality of a chatbot (e.g., performance, humanity, affect, and accessibility)[54]. Particularly relevant to current research aims are the performance and affect qualities. Conversational agents should pass the Turing test and include human errors to add an element of realism since communication between humans is generally not perfect. Interaction should be natural and seamless. Affect is a particularly appropriate quality for the current project given the intended purpose of assessing empathetic conversation among soldiers. Conversational agents should provide greetings, convey personality, warmth, and authenticity. They should to offer and evaluate emotional cues (read and respond to the state of the human user). The use of a quality emotion-based dialog corpus in conjunction with general conversation corpus (e.g., Cornell Movie Dialog Corpus [19]) will allow for the facilitation of affective qualities [54].

3.3 Overview

Overall, traditional conversational agents, generally do not produce human-level dialogue exchanges. The tasks like the Turing Test, continue to pose a challenge for even modern dialog systems. Thus, this task remains to be a hard problem of Artificial Intelligence. Still, recent advances allow for practical and realistic interaction modeling between human users and conversational agents. A platform like DialogFlow allow for cost-effective development of highly adaptable and flexible free-range dialog exchanges at a low cost, with a relatively small dataset. This may be because of the inner workings of DialogFlow; which may be parallel to state-of-the-art models that utilize the Seq2Seq architecture. Current research explores the use of both DialogFlow chatbot development and implementation of the Seq2Seq architecture.

Chapter 4

Deep Learning

The most common goal of artificial intelligence is to make machines adaptable [56]. Adaptability is a strength that has been commonly identified within human cognition (e.g., visual or audio perception). For this reason, machine learning experts may often seek to draw inspiration from the inner workings of neuro-anatomy. This source of inspiration facilitates the design of algorithms that may be used to tackle daunting computational tasks (e.g., speech or image recognition). Most traditional computer programs are not adaptable. Traditional computing facilitates quick arithmetic calculations and the efficient sequential execution of predefined instructions. However, when applied to problems that require human intelligence and computational flexibility, traditional computing has its limits. An example of such a problem is the automatic recognition of handwritten digits [56]. In what way could a program be written to capture the nuances of this type of tasks flexibly? Perhaps, a rule like, 'if it is a closed loop, the digit is a 0', could be considered, however, there would be many exceptions to consider (i.e., what if the loop is incomplete?). A program like this could be convoluted and lengthy. Biologically-inspired computational techniques offer a more sophisticated approach. Deep Learning is a field that has experienced numerous advances within the past ten years and may provide better ways to solve problems that require human intelligence. Current research aims to develop a computational approach for classifying and generating emotional text-based utterances. For problems that require a high level of human intuition, Seq2Seq deep learning architectures may be more effective [56].

4.1 Deep Learning, A Biologically Inspired Machine Learning Approach

Deep Learning is a field that may be challenging to define universally. However, there is one consensus, deep learning is based upon a set of experience-driven machine learning techniques. Machine Learning, a subset of Artificial Intelligence, provides techniques that allow models to learn through the presentation of examples. These techniques are reflective biological learning (i.e., changes in the way neurons respond to salient environmental stimuli, over time) [56].

Machine Learning algorithms are more concise than traditional rule-based programs. A small set of rules supplement Machine Learning models and allow for a decrease in the error during each iteration of training [56].

4.1.1 Recent Deep Learning Approaches to Natural Language Understanding

Experience-driven machine learning techniques are becoming a ubiquitous standard for approaching Natural Language Understanding (NLU) tasks (e.g., emotion extraction from the text). Two popular NLU algorithms include Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). Both sets of algorithms have produced high accuracy rates on emotion ([57]) and sentiment ([58]) analysis problems [59].

4.1.2 Approaches using CNNs

Convolutional Neural Networks (CNNs) extract high level features for word constitutions (n-grams) and have been used for a variety of linguistic tasks including, sentiment analysis [51] , text summarization[60], question answering systems [61], and sarcasm detection [62]. These hierarchical systems rely on computational techniques like max pooling to mimic biologically inspired sensation perceptual systems (e.g., visual information processing which relies on a hierarchy of specialized cells between light processing cells in the eye and the visual cortex). The strength of CNNs is the algorithm's ability to extract the most important word constitutions(n-grams) from text [63].

The Latin etymology of the word convolution, 'convolere,' means to 'roll together.' Convolutional Neural Networks are inspired by the visuosensory perception that is the hierarchical aggregation of lower level information (environmental stimuli) to higher level information. Similarly, word n-grams are aggregated into sentence vectors, as information passes through convolutional and max-pooling layers. For these reasons, CNNs are adept at high-level feature extraction and semantic representation but utilized within the current architecture due to the general inability of CNNs to capture long-range temporal dependencies [64].

4.1.3 Approaches using RNNs

RNNs offer an effective way to model the inherent sequential nature of language. The inner-workings of this algorithm will be further explored later. Additionally, RNNs perform robustly on the extraction of context dependencies from utterances of variable length. Multi-level text categorization [64], image captioning [65], speech recognition [66], machine translation [39], multi-modal sentiment analysis [67], subjectivity detection [68], sentiment analysis [69], emotion detection [70], and language modeling [71] are all tasks to which RNN approaches have been effectively applied. [72]

4.1.4 Training a Neural Network through Backpropagation

Backpropagation is the algorithm that facilitates the "learning" process within artificial neural networks. Within, traditional networks, the goal of backpropagation is to minimize training error (the difference between target output values and actual output values) through the iterative adjustment of weights within the network's layers [73].

The back-propagation algorithm within RNNs is slightly modified. Within RNNs the backpropagation algorithm is referred to as, Backpropagation Through Time (BPTT). The main difference is that this algorithm allows for training that "unfolds" temporally through time. In this way, weight updates are aggregated during each time step. [73].

4.1.5 RNN Cell Types

The basic unit of an RNN is the recurrent cell, which contains parameters within matrices. These matrices utilize the current input and current state of the cell to compute the upcoming state and output values.[72] The most basic representation of the recurrent memory cell is a simple RNN. The simple cell takes two input values (current and previously hidden states) and generates two output values. An output that can be passed to the next hidden layer or softmax layer is computed. The other output is fed back into the hidden state [72].

Within the simple RNN cells, inputs are concatenated together as they pass through the feed-forward layers. The computation of the current state involves the repeated multiplication of coefficients within the weight matrices and can, therefore, lead to a couple of well-known problems within neural network training: exploding gradient and vanishing gradient. When some of the coefficients with the weight matrices are greater than one output values may become very large too soon. When coefficients are close to zero, the network quickly loses information as the network training propagates through time. The solution to the exploding gradient within traditional natural networks is the use of a nonlinear function that limits the coefficient to one. In order to solve this problem, alternative RNN cell types have been developed(i.e., LSTM, GRU) [72]. These cell types will be discussed further in upcoming sections.

4.1.5.1 Long Short-Term Memory (LSTM)

LSTM cells originated in 1997 and are more complex than the traditional vanilla RNN. These cells have three gates (i.e., forget gate, input gate, and output gate) that limit the amount of information processed by the network at a given time step. The forget gate determines how much of the previously hidden gate to pass on to the next time step; the input gate determines how much the current input to consider in the computation of the output, and the output gate determines how much of the output will be included by the hidden state. This modified approach allows for increased management of long term dependencies. The LSTM architecture addresses the vanishing gradient problem through the use of an activation function that utilizes an identity activation function with a derivative of one. This way, the gradient neither subject to vanishing or exploding and remains constant throughout network training [72].

4.1.5.2 Gated Recurrent Units (GRU)

The Gated Recurrent Units (GRU) is an alternative cell type that addresses the vanishing and exploding grading descent problems seen within traditional RNNs architectures.[73] It does so through the use of two gates (i.e., reset gate, update gate). These gates incrementally determine how much of the hidden state's information is retained or forgotten throughout the sequential information processing steps. This architecture is simpler than that of the LSTM. Within this architecture, there is no need for an output gate. Thus, fewer parameters are required through the use of this cell type within the architecture. Furthermore, it utilizes a single state vector that does not distinguish between the current and hidden step. Recent comparative research has concluded that LSTM and GRUs both have superior performance compared to the vanilla architecture; however, there are no indicators that one performs better than the other. [73] [59]

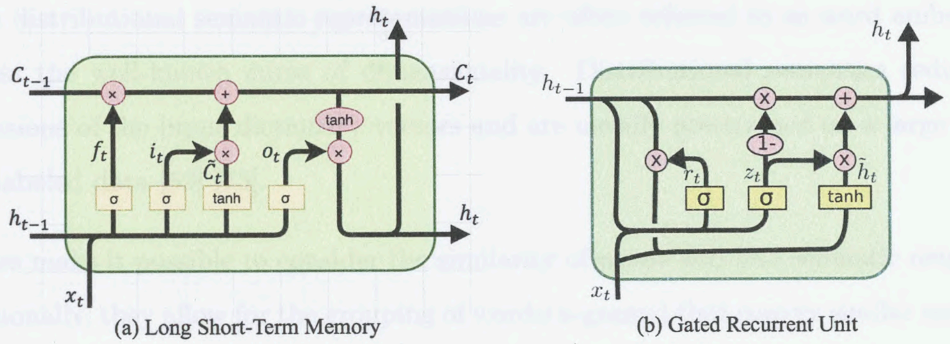


FIGURE 4.1: GRU and LSTM cell architecture [1]

4.2 Examining RNN Model Inputs: Word Embeddings

Now that a foundational understanding of the Recurrent Neural Network, the learning process, and the basic units that facilitate this learning (cells), has been established the model input may now be considered. Recurrent Neural Networks, like all ANNs, take in inputs as numerical values. For this reason, for inputted sentences, words must first be converted into numerical representations. Likewise, characters may be converted into numerical representations, depending on the needs of the algorithm. The one hot word vector is the most simple and naive input form. It consists of an array that is the same length as the input dictionary. Within this representation, indices that do not correspond to a word in the dictionary contain zeros [73].

Other approaches combine stemming and lemmatization. Stemming reduces the size of the vocabulary by stripping words down to just their roots (without the suffixes) while lemmatization applies more powerful rules for the removal of information within the vocabulary [73].

All of these approaches have one huge limitation. There is no consideration of semantic and syntactical information or the relation between words. These traditional embeddings leave the model to do all of the arduous work of extracting meaning. The inclusion of prepossessing layers at the beginning of the neural network that predicts words based on their context allows for training that is quicker and more efficient. Distributional semantics address this weakness through mapping of words a dense real vector of fixed dimensions. This optimized semantic distributions approach allows for the retention of the maximum amount of relational information [73].

These distributional semantic representations are often referred to as word embedding, address the well-known curse of dimensionality. Distributional semantics reduce the dimensions of the input dictionary vectors and are usually pre-trained on a large corpus of unlabeled data [59] [73].

Vectors make it possible to consider the similarity of words and find semantic neighbors. Additionally, they allow for the grouping of words(n-grams) that convey similar meanings (e.g., n-grams)[59] [73].

Word embeddings were highly popularized by works that made use of the Continuous Bag of Words (CBOW) and Skip-Gram Models for the production of high quality distributed vector representations [63] [73].

4.2.1 Word2Vec Embeddings

Mikolov [74] revolutionized word embeddings with a novel contextuality based approach. This approach utilizes Continuous Bag of Words (CBOW) models and Skip Gram models. CBOW models compute the conditional probability of the target words based upon the context of words in a surrounding window of a specified size, while skip grams do the opposite. They predict the surrounding context of words given a central target word. This approach is particularly useful because it allows for the representation of word groups (phrases). However, one concern is that, depending on the size of the context windows, word of different sentiments (i.e., good, bad) can be labeled with the same embedding [63] [73].

4.2.2 Character Embedding

While word embeddings are adept at capturing syntactic and semantic information they may not be adequate for some other Natural Language Understanding(NLU) tasks like POS Tagging, Named Entity Recognition (NER), and intra-word morphology determination. NLU systems that process character-level input may be useful for such tasks. Character-Level embeddings have been particularly effective on multilingual neural language models. Making them particularly useful is their ability to deal effectively with Out of Vocabulary (OOV) words (e.g., grammatical errors) [59].

4.3 RNNs Augmented with Attention

RNN model-augmentations like the use of LSTM cells and distributed representations (word embeddings) have improved the preservation of information within long-term context dependencies. However, this model has been further improved through the use of attention mechanism techniques [63] [73].

The traditional encoder-decoder architectures have been required to process irrelevant information for the model task. The processing of the irrelevant information has been prone to decreasing overall model efficiency, especially when lengthy or information-heavy inputs are supplied. One way this limitation has been mitigated is through the use of selective encoding techniques that allow for the continual back-referencing to the encoder. A context vector is calculated using an input hidden state sequence is utilized by the decoder as a way of keeping a reference of only the most pertinent information for task-handling within the encoder. Namely, this particular technique is known as an, attention mechanism. Sequential Tasks like text summarization and machine translation (e.g., [39], [75]) have been improved through the use of attention mechanism techniques [63] [73].

4.4 More on Tackling Sequential Machine Learning Tasks with RNNs: Seq2Seq Architecture

Sutskever, Vinyals, and Le [1] examined the development of a model that allows for the mapping of sequential information to vector representations, that is sequences to sequences. A multi-layered LSTM that mapped inputted vectors to a fixed-dimensionality vector was utilized in conjunction with another LSTM that decided the sequence from the context vector. State of the art results was achieved (i.e., BLEU score of 34.81 on Neural Machine Translation task). Since the introduction of this new technique, this model has become the state of the art on sequential information processing research. This architecture is generally known as Seq2Seq and will be utilized for the modeling of sequential information with affective-conversation within specific conversational domains (scenarios) for the current research aim [1].

As previously mentioned RNNs are adept at mapping sequences to sequences particularly when alignments between the inputs and outputs predetermined. Weaknesses of the vanilla RNNs were previously examined. Particularly those weaknesses include long-range temporal dependencies limitations and the loss of information as propagation through network takes place. The seq2seq architecture is the specific model that addresses the deficiencies through the use of the combined LSTM encoder and LSTM decoder networks [1].

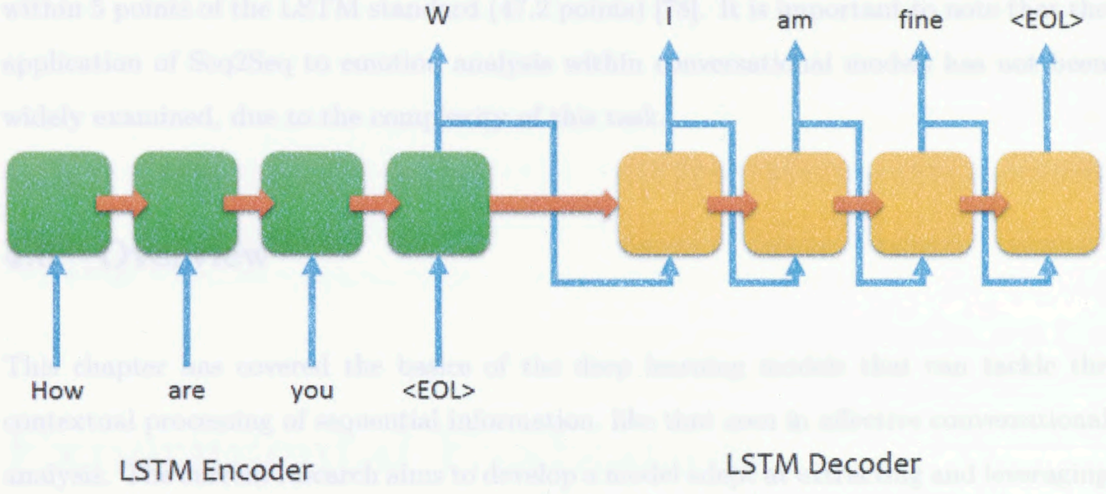


FIGURE 4.2: As seen, here given the source sequence x1, x2,. . . .xm and a target sequence y1, y2, ...ym this model estimates the conditional probability Pr(y1, y2, ...yn | x1, x2,. . . xm). As seen in the figure, "How", "are", "you", "<EOL>" computes the probability of the sequence "I", "am", fine", "<EOL>" where EOL represents a special end-of-sentence symbol. [1] [76]

4.5 Applying Seq2Seq Architecture to Textual Affect Analysis

Long-term temporal context dependencies allow for state of the art of processing both long and short term dependencies. This has been shown in previous bodies of literature that have examined the application of the architecture to sentiment analysis tasks and some aspects of textual affect analysis (e.g., [77]). Making this architecture particularly useful, is the maintenance of input representations for utterances of variable length. Furthermore, this model allows for the consideration of relevant and appropriate information for the output of responses, through the use of attention mechanism-based model embeddings that focus on selective regions of an inputted sentence). This means that from a human perspective, the model output should make more sense to a human user given the intent of the inputted emotional utterance (e.g., input: "SGT you've done a fantastic job with keeping the platoon in top shape. Great work!", response: "Thanks Sergeant I am elated to hear that."; vs input: "SGT you've done a fantastic job with keeping the platoon in top shape. Great work!", "I know what I am doing. Stay out of my business and get your platoon in shape.") Furthermore, based on previous research current predictions of model support the prevalence of an evaluative perplexity score within 5 points of the LSTM standard (47.2 points) [78]. It is important to note that the application of Seq2Seq to emotion analysis within conversational models has not been widely examined, due to the complexity of this task.

4.6 Overview

This chapter has covered the basics of the deep learning models that can tackle the contextual processing of sequential information, like that seen in affective conversational analysis. The current research aims to develop a model adept at extracting and leveraging emotional data from text. RNN models, particularly the Seq2Seq architecture, facilitates robust sequence analysis tasks, and allow for the training of model's that can "remember" previously inputted emotional cues. [1] Leveraging the strength of these models will be pertinent to current and future research endeavors. The upcoming chapter will cover experimental evaluations of the currently proposed model, using two primary platforms, DialogFlow and a TensorFlow based implementation in Python. Both of these approaches offer ways to extract emotionally and contextually relevant information from text within conversational agents, utilizing the seq2seq architecture.

Chapter 5

Experimental Evaluation

This chapter presents the evaluation of two comparative methodologies for two machine learning tasks: emotion classification, and emotional response generation. Methodologies for the first task consists of the following: the development of a conversational dialogue corpus (i.e., Real World Professional Conflicts Dialog Corpus), the training of a Dialog Flow-based emotion classifier with the compiled dataset. The Real World Professional Conflicts Dialogue Corpus contains utterances in four emotional regions on the Circumplex Model of Affect. The second task entails two phases of model training: 1) training a generative conversation model on a general conversation corpus (i.e., Cornell Movie Dialogue Corpus); 2) Fine-tuning the conversational model by training on a specific set of emotion-based utterances (i.e., Real World Professional Conflicts Dialog Corpus). Upcoming sections will further detail the experimental evaluation of the current research aim.

5.1 Evaluation Metrics

The classification task is performed utilizing the DialogFlow conversational agent platform. This platform allows for the training of various chatbots utilizing intent recognition and entity. Accuracy is the primary metric utilized in evaluating the performance of this model.

Evaluation of the generative conversational model makes a note of the following definition of conversational modeling: “The mapping of utterances to responses[79]. The current research aim utilizes a combined encoder LSTM (which takes in the utterance) and decoder LSTM (which produces a response to that utterance). Thus the general aim of current methodological evaluation is to observe how well the generative model maps inputted responses to target utterances within the intended emotional and conversational contexts. This is a highly complex task to evaluate. Currently, there are no automatic evaluation metrics for chatbots. Thus, many chatbots are evaluated with a Turing Test. Turing Test evaluations can be costly and time-consuming. For this reason, automatic neural machine translation metrics may be more useful for the current evaluation process. Many traditional seq2seq conversation tasks have done so utilizing the neural machine translation automatic evaluation metric, BLEU (i.e., inspired by Google's zero-shot multilingual translation system [80]). The general goal of BLEU is to determine how well the model produces reasonable responses to input (i.e., carry out decent conversation and the level of entropy among inputted and outputted utterances) [27] [79].

5.2 Datasets

5.2.0.1 Real World Professional Conflicts scenario-Corpus

The Real World Professional Conflicts Corpus was developed at Army Research Institute facilities at the Fort Benning Installation. Over 3400 utterances were collected in response to Army-specific real-world professional conflict scenarios. The majority of utterances were collected at MCoE Army Training facilities while the remaining utterances were collected via a Qualtrics delivered assessment. Some SONA participants received varying degrees of extra credit as determined by various instructors in the Psychology department.

A self-report survey was administered through Qualtrics, a web-based survey deployment platform available through Columbus State University's Technology Department [81]. The sample included participants who were 18 years of age or older from all demographics. Approximately 49 participants from the Columbus State University Psychology Department provided responses that contributed to this corpus. Responses for 29 of these participants were utilized in the dataset. The remainder of these responses were set aside for future research endeavors at ARI facilities. These responses were not utilized in current iterations of training, because these utterances did not correspond to the four main domains(scenarios) of interest; or were incomplete. A total of 96 students from IBOLC and ABOLC provided responses that contributed to this corpus. A free response battery was delivered to junior officers in MCoE facilities via Portable Document Format-based assessments. To optimize the quality of the dataset, responses from participants who did not respond conversationally, and instead responded in an instructional manner (e.g., "Tell the soldier that he needs to straighten up his attitude or he's out") were omitted. Participants were asked to provide two responses that corresponded to each of the four quadrants on the Circumplex Model of Affect [4]. The responses were generally limited to 255 characters.

There were two sets of scenarios with overlapping content created for the current research aim. The first consisted of 12 scenarios and the second consisted of 4 scenarios. Within each of these sets, scenario descriptions were between 100 and 250 words long. For the first set eight of the 12 scenarios were domain specific (i.e., Army Training-Based Scenarios). Four of the 12 scenarios were domain-nonspecific (civilian-related professional conflicts). Here is a general description of each these scenarios:

- Scenario 1 (Absenteeism): Presents a situation involving a worker at a law firm who calls out sick from work for four days and forgets to thank his coworkers for taking over his responsibilities, during his hiatus. [Source: RealCareer™ Employability Skills Program Scenario Resources.] [82]
- Scenario 2 (Respect): Presents a situation involving a grocery store clerk who arrives to work in an irritable mood and is seen yelling at one of his coworkers out this frustration. [Source: RealCareer™ Employability Skills Program Scenario Resources.] [82]
- Scenario 3 (Depressed Staff Writer): Presents a situation involving a staff writer at an established magazine company who feels overwhelmed with external life stressors (e.g., a child who is ill, the financial strain of medical bills, financial undercompensation at work). This character reveals that she is abusing depression medications during a counseling session for inconsistent work performance.
- Scenario 4 (Patience): Presents a situation involving a team member who shows up late to a major conference presentation, unapologetically. [Source: RealCareer™ Employability Skills Program Scenario Resources] [82]

- Scenario 5 (ELDMIC Motor Pool): Presents a situation involving the delay of a field training exercise development due to systemic training issues. Despite the collective frustrations of the overall platoon, one subordinate is openly expressing her elation about not having to complete the required exercise. [Source: *Designing Scalable, Objective Assessments of Interpersonal Leadership Skills*; *Building Automated Assessments of Interpersonal Leadership Skills*][5] [83]

- Scenario 6 (Taking Charge): Meeting the Platoon Sergeant): Presents a situation involving the exchange between a new platoon leader and an Acting Platoon Leader as the former leader reveals details regarding a potential problem in the state of the platoon. The acting platoon leader is providing an overview of the state of the platoon and is expressing concerns for a squad leader who struggles with alcoholism. [Source: *Videodisc Interpersonal Skills Training and Assessment (VISTA). Volume 3. Scenarios*] [84]

- Scenario 7 (Verbal Abuse): Presents a situation involving counseling of a Squad Leader who has repeatedly been verbally abusing his squad members due to marital problems that he has been facing at home. [Source: *Videodisc Interpersonal Skills Training and Assessment (VISTA). Volume 3. Scenarios*] [84]

- Scenario 8 (Hand Receipt Altercation Interjection): Presents a situation involving a physical altercation between two subordinates. [Source: *Designing Scalable, Objective Assessments of Interpersonal Leadership Skills*; *Building Automated Assessments of Interpersonal Leadership Skills*][?] [83]

- Scenario 9 (Emergency Crisis - Suicide Threat) Presents a situation in which a subordinate expresses suicidal ideation after being called in for counseling for inconsistent work performance. [Source: *Videodisc Interpersonal Skills Training and Assessment (VISTA). Volume 3. Scenarios*] [84]

- Scenario 10 (Emergency Crisis - Emergency Leave) Presents a situation in which a Staff Duty Officer requests to go home to attend to the acute, fatal illness of a family member. [Source: *Videodisc Interpersonal Skills Training and Assessment (VISTA). Volume 3. Scenarios*] [84]

- Scenario 11 (Insubordination - Haircut): Presents a situation involving a subordinate who is refusing to get his haircut to military standards and is defiant of disciplinary procedures. [Source: *Videodisc Interpersonal Skills Training and Assessment (VISTA). Volume 3. Scenarios*] [84]

- Scenario 12 (Performance Counseling): Presents a situation in which a section leader who is being counseled for his erratic performance and shows up unapologetically late to his counseling session. [Source: *Videodisc Interpersonal Skills Training and Assessment (VISTA). Volume 3. Scenarios*] [84]

Initially, all 12 of these scenarios were utilized during the initial phases of the data collection (see Appendix A). However, these were reduced down to four specific Army related scenarios, for simplified analysis. A record was kept with the initial responses for the original set of scenarios. However, they were not utilized within the scope of the current research aim. These utterances may instead be utilized for future research endeavors at ARI facilities. Thus, the finalized Real World Professional Conflict scenarios Corpus consists of responses to only four of the original scenarios (i.e., Scenario 6[Taking Charge], Scenario 7[Verbal Abuse], scenario 9 [Emergency Crisis - Suicide Threat], Scenario 12 [Performance Counseling]).

Participants were presented with scenarios in which they took on the role of leader (i.e., Lieutenant) who is conversing with a professional subordinate after the presentation of a conflict.

Within each scenario, the target characters initial emotional state begins in a different emotional region (see Appendix A). As previously mentioned, participants were asked to generate two responses that would transfer the character's emotions from one state to another within the four quadrants on the Emotional Circumplex Model of Affect [4]. The four quadrants included the following regions. Region 1: Contains states that are higher in emotional intensity (activation) and more highly pleasant (higher valence). Region 2: Contains states that are higher in emotional intensity and less pleasant (lower valence). Region 3: Contains states that are lower in emotional intensity (activation) and less pleasant (lower valance). Region 4 Contains states that are lower in emotional intensity (activation) and more highly pleasant (higher valence).

Examples of emotions in Region 1 includes states that are happy, enthusiastic, excited, elated, and alert. Examples of emotions in Region 2 include states that are frustrated stressed, nervous, intense, upset, and angry. Examples of emotions in Region 3 include states that are sad, depressed, bored, lethargic, and fatigued. Examples of emotions in Region 4 include states that are relaxed, complacent, contented, serene, calm, and comforted.

The final dataset, the Real World Professional Conflicts scenario-Corpus, was utilized in the training of classification and generative conversational model.

5.2.1 Cornell Movie-Dialogs Corpus

This dataset was developed by Cristian Danescu-Niculescu-Mizil and Lilian Lee and contained over 220,579 conversational exchanges (i.e., 9,035 characters, 617 movies, 304,713 total utterances) [19]. The conversations within this corpus are not realistic to everyday professional dialog (i.e., themes include love, violence, murder). While it would be more appropriate to utilize a dataset with containing conversations within a professional domain; this dataset is well formatted and allowed for Phase I model training. Generally, the use of this dataset produced an erratic conversational flow. This weakness may be addressed through the conduction of two separate training phases which will be discussed in an upcoming section [19].

5.3 Methodology

The methodology is specific to each conversational task covered under the scope of the current research aim: that is the classification conversation modeling task using DialogFlow generative conversational modeling task using TensorFlow, an open-source machine learning library [85][86].

5.3.1 Classification Task Methodology

DialogFlow [86] was used for the training and evaluation of the classification model. Participants provided responses that were used to train four intents corresponding to the four main regions on the Circumplex Model of Affect. Approximately 200 utterances were used to train each intent. Overall, over 800 utterances were utilized for the training and testing of each model(90 percent training [e.g., approximately 720 utterances], 10 percent testing [e.g., approximately 80 utterances for each scenario]). The further analysis evaluated model performance at different training to testing data ratios for two of the four chatbots (e.g., 70 percent training, 30 percent testing; 80 percent training, 20 percent testing).

Overall, a methodology for the classification task entails of the following steps.

1. A dataset of 3400 utterances was divided into four datasets. Each of these four datasets contained over 800 utterances corresponding to four primary emotion regions on the Circumplex Model of Affect and was presented during the data collection process of the Real World Professional Conflicts Dialogue Corpus.

2. Each of the four datasets was split again into four smaller datasets corresponding to the four primary emotion regions (e.g., Region 1 [high valence, high intensity], Region 2 [low valence, high intensity], Region 3 [low valence, low intensity], Region 4 [high valence, low intensity]. Each dataset contains over 200 utterances specific to a corresponding emotional quadrant.

3. Four chatbots, SFC Johnson, SSG Burch, PFC Lewis, SSG Rogers, were created using DialogFlow; the Machine Learning Classification Threshold was set to a confidence score of 0.3. Thus, inputted dialog corresponding with classification confidence scores less than 30 percent were triggered the fallback intent. Four user-specified intents, corresponding to the four primary emotional regions were created.

4. Model training was conducted for each of the four chatbots. A CSV containing ninety percent of the subsetted data points was uploaded for training for each of the four emotion-based intents. The remaining 10 percent of utterances, within each subset, was set aside for model testing. Sixteen total CSV files were uploaded, and 16 intents (among the four chatbots) were trained to utilize the uploaded data points.

5. For automated intent detection a webhook, an HTTP request, was enabled within each of the four chatbots. Ngrok, a web-tunneling tool was employed and allowed for the exchange of JSON formatted data between the client (a local Python-based micro web framework, Flask,) and server (Google's DialogFlow API). More details are included in Appendix B.

6. Using the python-client based webhook, model testing was conducted for each chatbot. The testing set consisted of 10 percent of the sub-dataset corresponding to each intent.

7. Model testing was conducted for each of the four chatbots.

5.4 Generative Conversational Model Methodology

5.4.1 Sequence2Sequence(Seq2Seq) Model

The model is based on Google's Neural Machine Translation model (Vinyals, 2016). This model is a sequence to sequence architecture with attention mechanism. This mechanism allows the decoder to access hidden states of the encoder more flexibly, to leverage improved model predictions.

As mentioned, in Chapter 1, the algorithmic implementation of the generative conversation models was adapted from code sources that are inspired by the following Udemy Tutorial: Deep Learning and NLP A-Z^{TM}: How to create a ChatBot. The source code is documented and cited in Appendix D. For the current research, several customized modifications to the source code to support the generation of affective dialogue [6].

The Recurrent Neural Network (RNN) utilizes LSTM cells. The hidden units are set to a size of 512. The learning rate is set to 0.0001. The vocabulary threshold (minimum number of token occurrences) is set to five.

Eighty-three percent (150 utterances) of the modified Real World Professional Conflicts Dialogue was used for model training, and the remaining seventeen percent (30 utterances) were utilized for model evaluation.

The model utilizes a greedy approach, selecting the token with the highest conditional probability at each decoder step.

5.4.2 Training

Training for this model is inspired by the architecture of previous seminal research projects (e.g., Nguyen, Morales, and Chin [79], Huang et al., [27]). Due to the small number of training utterances, four chatbot models are trained over two phases. These four models correspond to the four regions on the Circumplex Model of Affect. Phase I trains the model on the Cornell Movie Dialog Corpus for 100 iterations. This training phase captures the general conversational flow. Phase II limits training to the domain-specific affective datasets (i.e., a modified subset of the Real World Professional Conflicts Dialog Corpus [for which response utterances are created by researchers]). This training phase limits training to only utterances specific to the emotional class (along with the Circumplex Model of Affect). Phase II allows the model to produce realistic expressions within an emotional, contextual domain.

Phase II allows the model to generate conversation more specific to the target domain, Army-interpersonal dialogue. Training is extended utilizing the Real World Professional Conflict Scenarios Corpus utterances specific to each emotional domain.

5.4.2.1 Vocabulary and Hyperparameters

The vocabulary of both datasets was combined (i.e., Real World Professional Conflicts Corpus). For the Cornell Dataset, tokens that appeared at least five times were included in the vocabulary. Training buckets were set at the following parameters: BUCKETS = [(10, 15), (15, 25), (25, 45),(45, 60), (60,100)]. Within this bucket, pairs of utterances with similar encoder-decoder lengths are grouped for training purposes. During training preceding tokens are used to predict upcoming tokens.

During the second phase of training (on the emotion-specific dataset), buckets were reduced to the following parameters: BUCKETS = [(50, 30)]. For this phase, buckets were reduced to one single, encoder-decoder pair, due to the decreased size of the training dataset.

5.5 Results

5.5.1 Affective Classification Model

Accuracy rates for emotion-based intent recognition among the four developed chatbots are presented within the confusion matrices in Figures 5.2, 5.3, 5.4, and 5.5. The overall performance for the developed DialogFlow chatbots was modest, yet promising (e.g., SGT Burch chatbot accuracy of 56.8 percent; Staff Sergeant (SSG) Rogers chatbot accuracy of 51.25 percent; Private First Class (PFC) Lewis chatbot accuracy of 48.75 percent). The overall accuracy rates for the Sergeant First Class (SFC) Johnson chatbot were lower than expected (i.e., identified the correct intent for 35.7 percent of evaluation utterances). A couple of factors may explain modest performance: a relatively low dataset and confusion of the given assessment task among participants. While most Seq2Seq models and evaluated using large datasets (i.e., thousands or millions of training utterances), this dataset was testing using over 800 utterances per chatbot, and 200 utterances per regional intent. The small data-set may influence the robustness of training within each regional intent the Circumplex Model of Affect, for this reason, the system may not recognize or adequately classify intents with infrequent utterances, "You're fired, Mate." This weakness may be mitigated through the use of a larger higher quality general conversation dataset that can improve the training of the smaller dataset through transfer learning. Another general issue was the confusion of the target task. Some participants provided responses that described a general response to the scenario instead of a specific response (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 cues.

A DialogFlow API (see Appendix B) was utilized as an evaluative classifier for testing utterances. Each of the four models was tested utilizing the testing utterances, and the resulting emotional classification was determined. A DialogFlow evaluation classifier was trained using the Real World Conflicts Dialog Corpus. Utterances (produced by the models) were passed to the DialogFlow server via a combined Flask/Ngrok application, during the evaluation phase. A resulting classification of the utterances was outputted via this API and used to evaluate the performance of the classification model.

DialogFlow chatbots were deployed through a Slack Integration. Figure 5.1 presents an example of a conversational flow between a human user and a character-based slack bot.

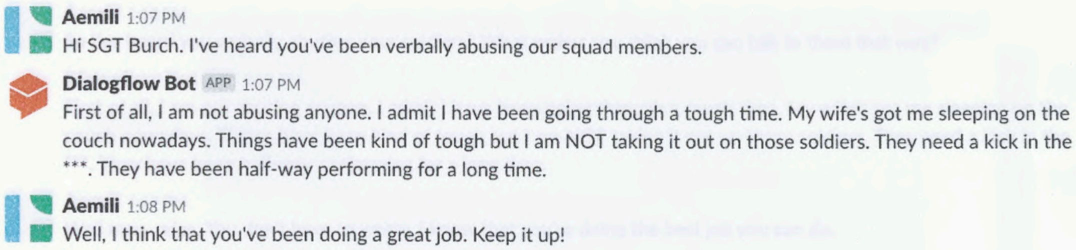


FIGURE 5.1: Sample conversation with DialogFlow chatbot, SGT Burch. SGT Burch's initial state is angry. In this conversational context, he is being counseled for verbally abusing his subordinates. Here, various emotional utterances with a high intensity-level and high valence-level invoke a similar emotional response from the chatbot.

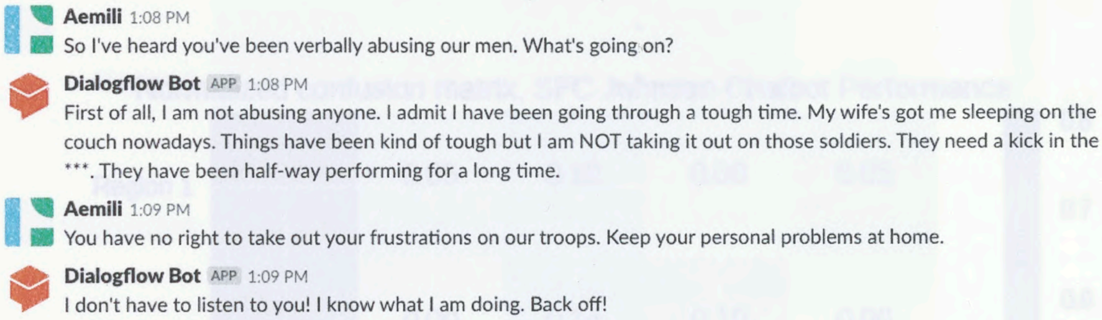


FIGURE 5.2: Sample conversation with DialogFlow chatbot, SGT Burch. SGT Burch's initial state is angry. In this conversational context, he is being counseled for verbally abusing his subordinates. Here, emotional utterances with a high intensity-level and low valence-level invoke a similar emotional response from the chatbot.

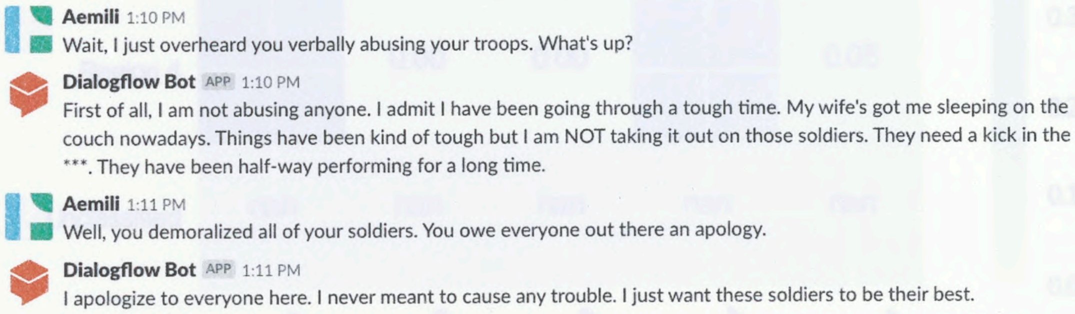


FIGURE 5.3: Sample conversation with dialog flow chatbot, SGT Burch. SGT Burch's initial state is angry. In this conversational context, he is being counseled for verbally abusing his subordinates. Here, emotional utterances with a low intensity-level and low valence-level invoke a similar emotional response from the chatbot.

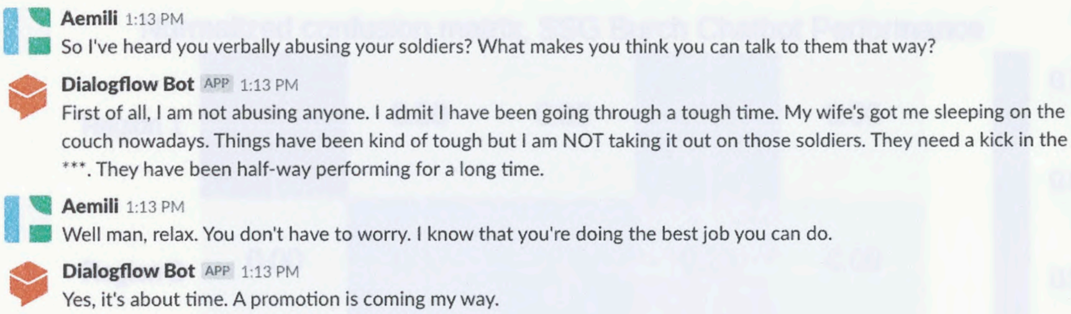


FIGURE 5.4: Sample conversation with dialog flow chatbot, SGT Burch. SGT Burch's initial state is angry. In this conversational context, he is being counseled for verbally abusing his subordinates. Here, emotional utterances with a low intensity-level and high valence-level invoke a similar emotional response from the chatbot.

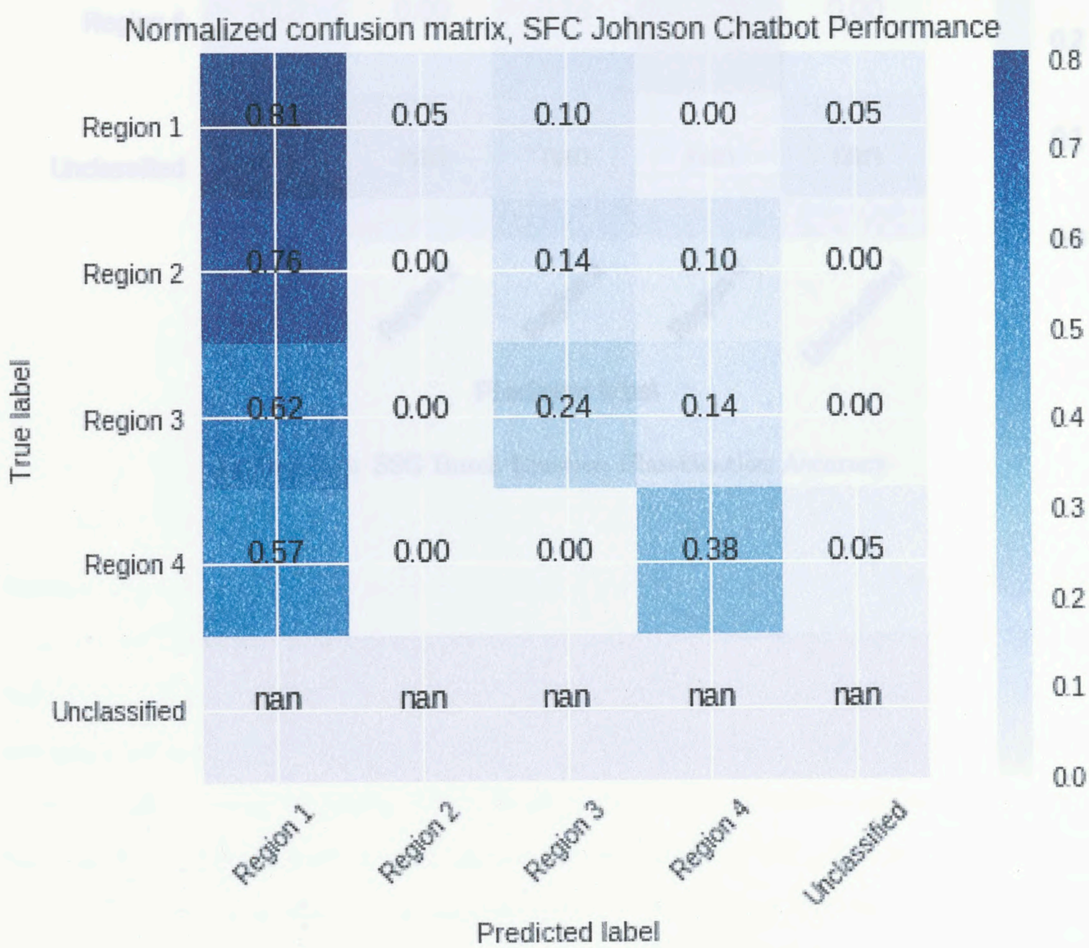


FIGURE 5.5: SFC Johnson Emotion Classification Accuracy

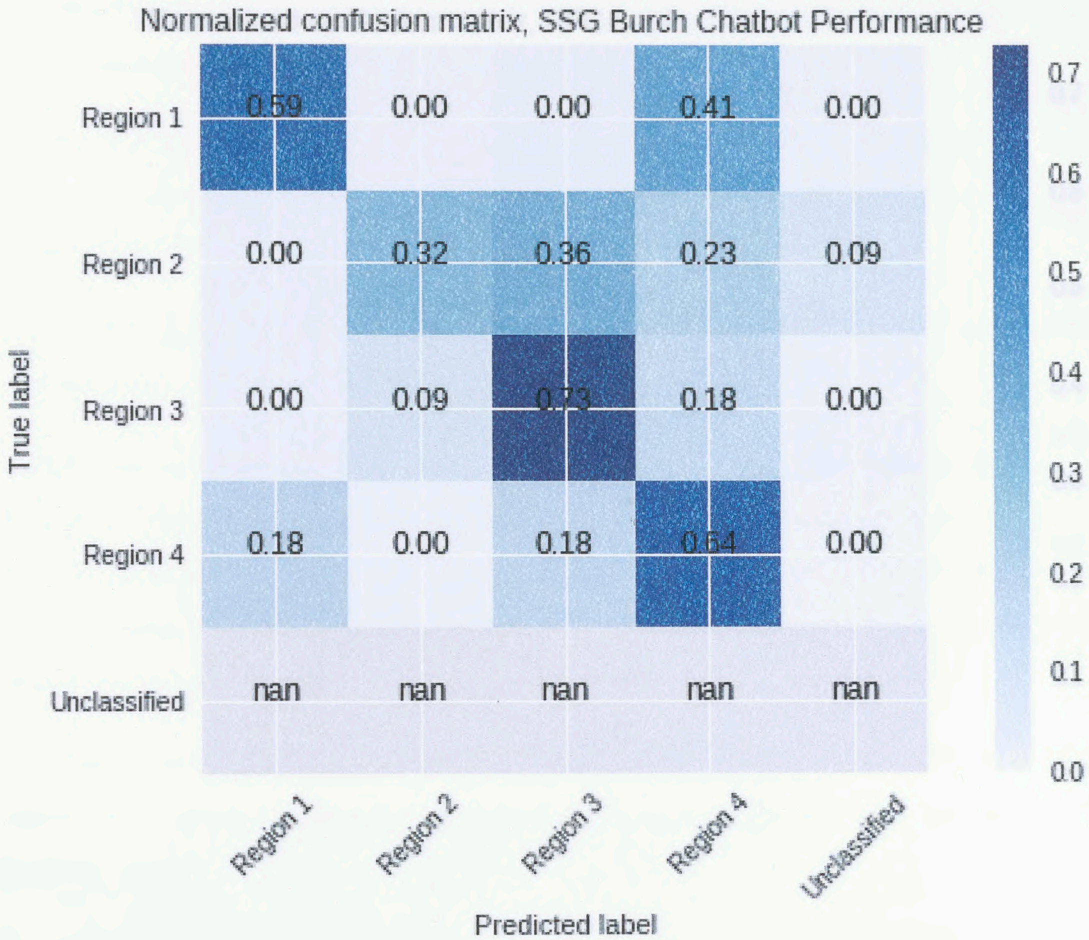


FIGURE 5.6: SSG Burch Emotion Classification Accuracy

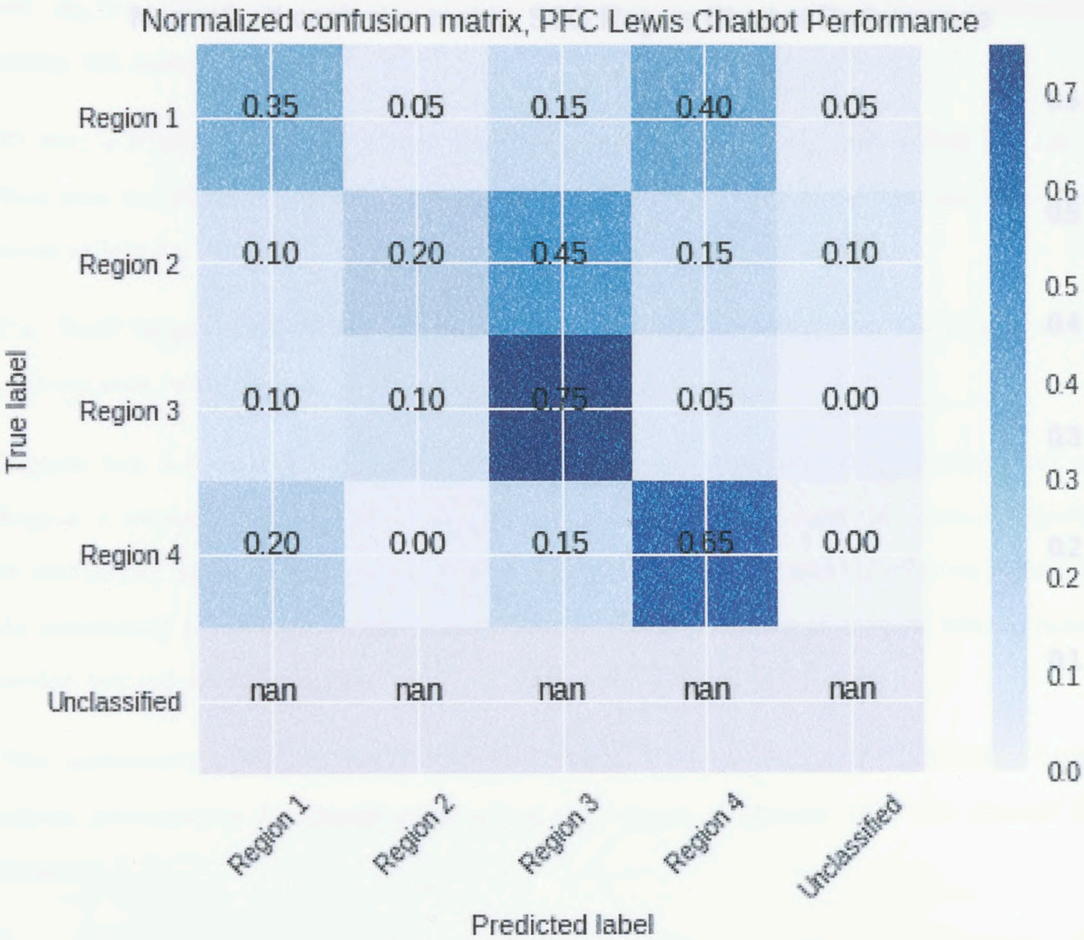


FIGURE 5.7: PFC Lewis Emotion Classification Accuracy

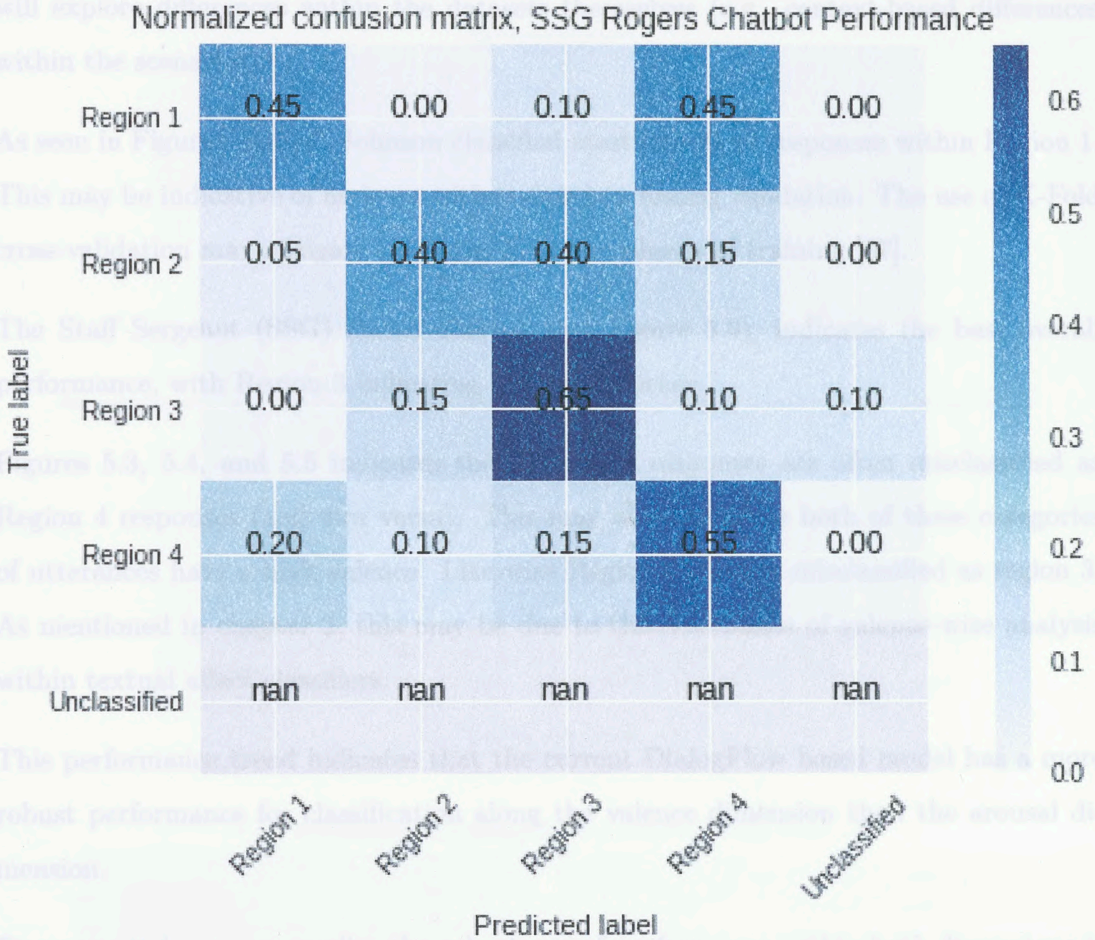


FIGURE 5.8: SSG Rogers Emotion Classification Accuracy

Figures 5.6, 5.7, 5.8 and 5.9 present the classification accuracy within two selected chatbots at three different training to testing ratios (i.e., 90 percent: 10 percent; 80 percent: 20 percent; 70 percent: 30 percent, respectively). As shown in the figures below, the classification accuracy generally improved with the increase in the training data. The chatbot corresponding to the character, SSG Burch, improved with each graduated increase in training data. The chatbot corresponding to the character SFC Johnson demonstrated counter-intuitive performance results. While SGT Burch's performance increases as the percentage of training data increases, SFC Johnson's performance generally decreases as the amount of training data increases. A possible explanation for unexpected performance trends could be the need for K-fold cross-validation [87]. K-fold cross-validation randomly systematically shuffles the data points. K-fold cross-validation may ensure that the test cases are not mainly edge cases within the dataset. Possible explanations for discrepancies in performance will be explored in further analyses. Future research aims will explore differences within the datasets themselves (e.g., context-based differences within the scenarios).

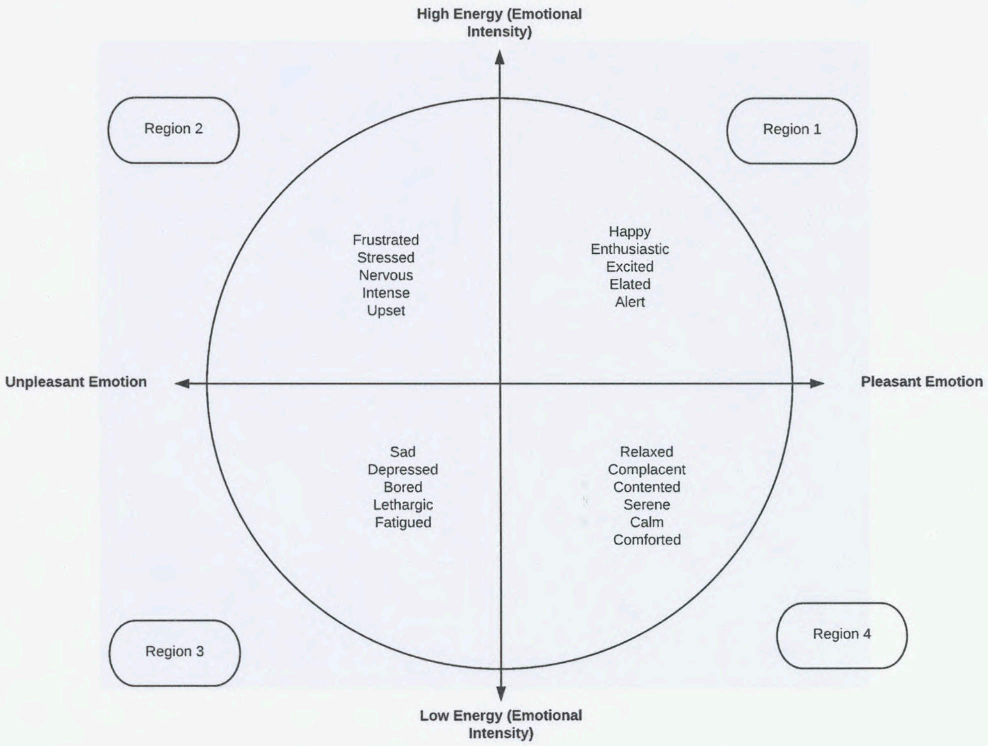
As seen in Figure 5.2, SFC Johnson classified most intents as responses within Region 1. This may be indicative of an issue with training to testing validation. The use of K-Fold cross-validation may mitigate this issue in future phases of training [87].

The Staff Sergeant (SSG) Burch chatbot (see Figure 5.3), indicates the best overall performance, with Region 3 indicating highest accuracy.

Figures 5.3, 5.4, and 5.5 indicates that Region 1 responses are often misclassified as Region 4 responses (and vice versa). This may occur because both of these categories of utterances have a high valence. Like-wise Region 2 is often misclassified as region 3. As mentioned in chapter 2, this may be due to the dominance of valence-wise analysis within textual affect classifiers.

This performance trend indicates that the current DialogFlow based model has a more robust performance for classification along the valence dimension than the arousal dimension.

One potential way, to equalize the robustness of performance within both dimensions is to develop a custom made-emotion classifier. While using a chatbot development tool like DialogFlow may be cost and time effective, users do not have access to internal hyper-parameters. Thus, fine-tuning model performance, beyond intent detection and entity extraction are challenging. Furthermore, training is less automated and may be laborious on the development side.



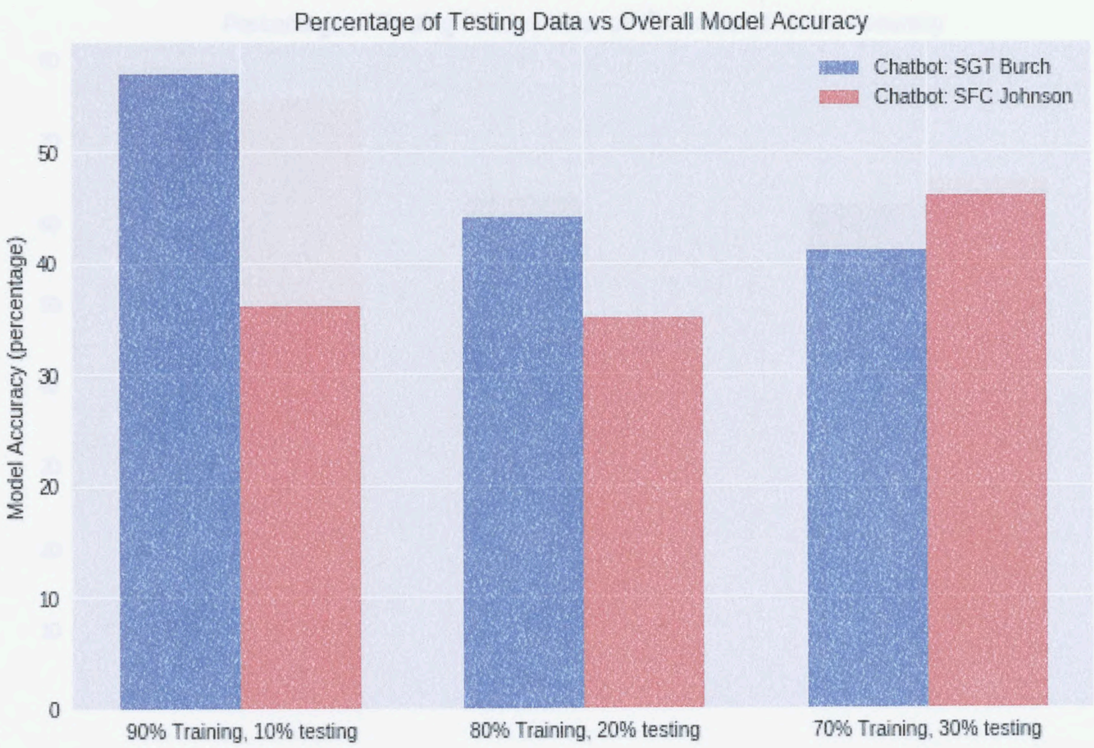
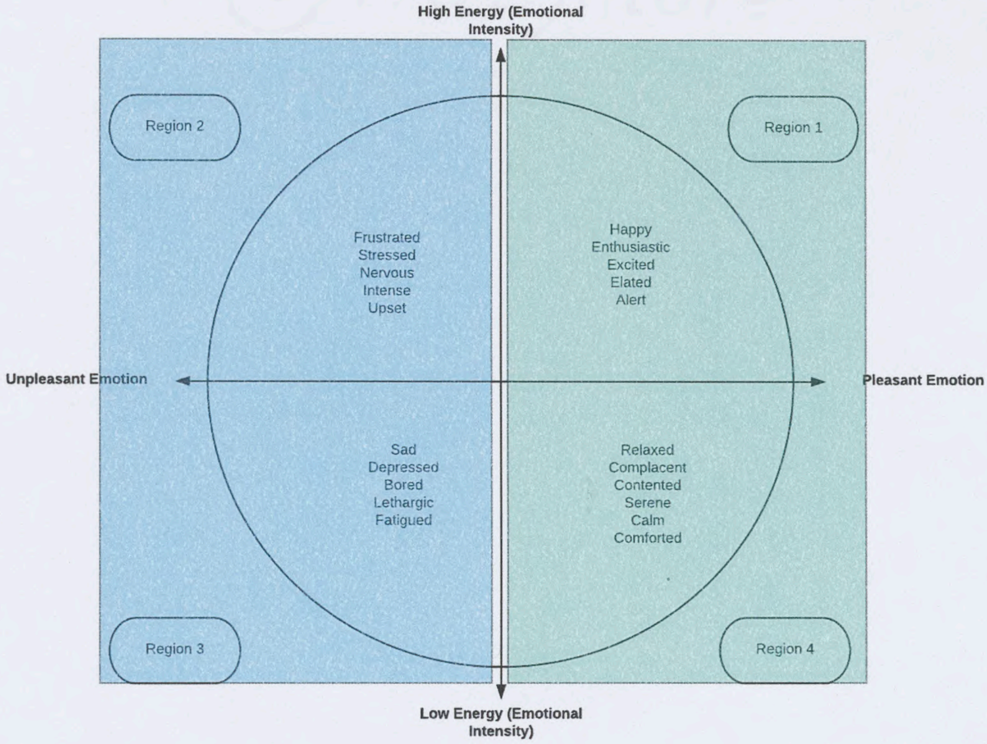


FIGURE 5.9: Overall classification accuracy at three comparative training to testing ratios for SSG Burch and SFC Johnson Chatbots. This dimension represents the classification accuracy among all four regions regions on the Circumplex Model of Affect.



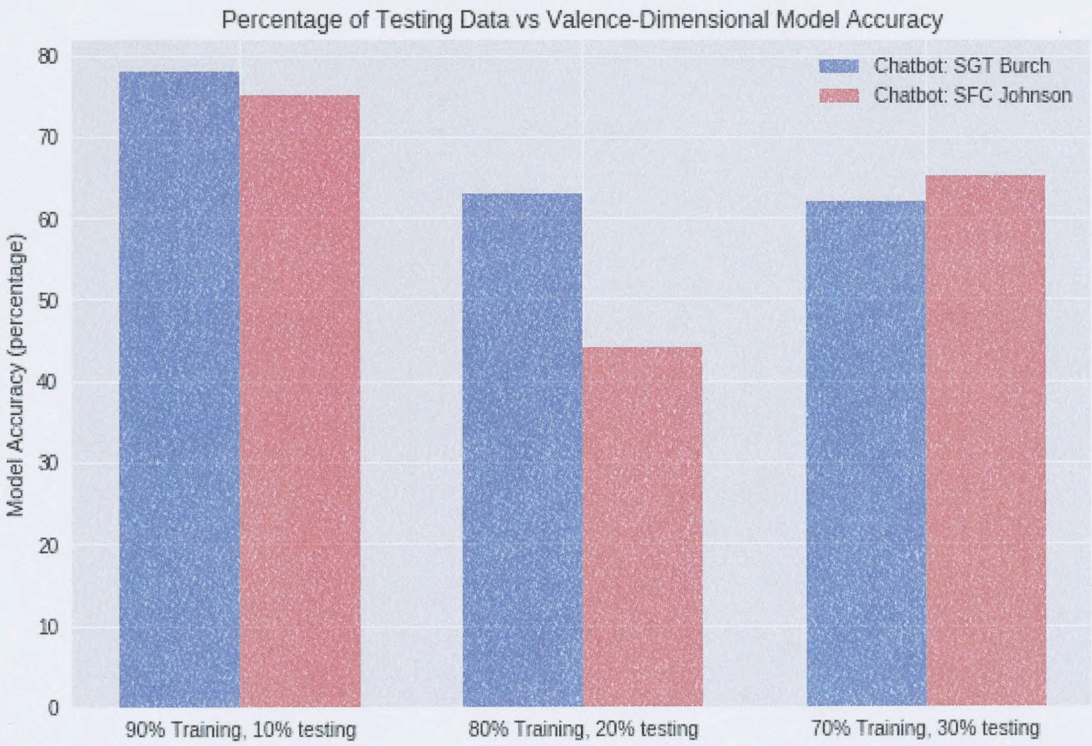
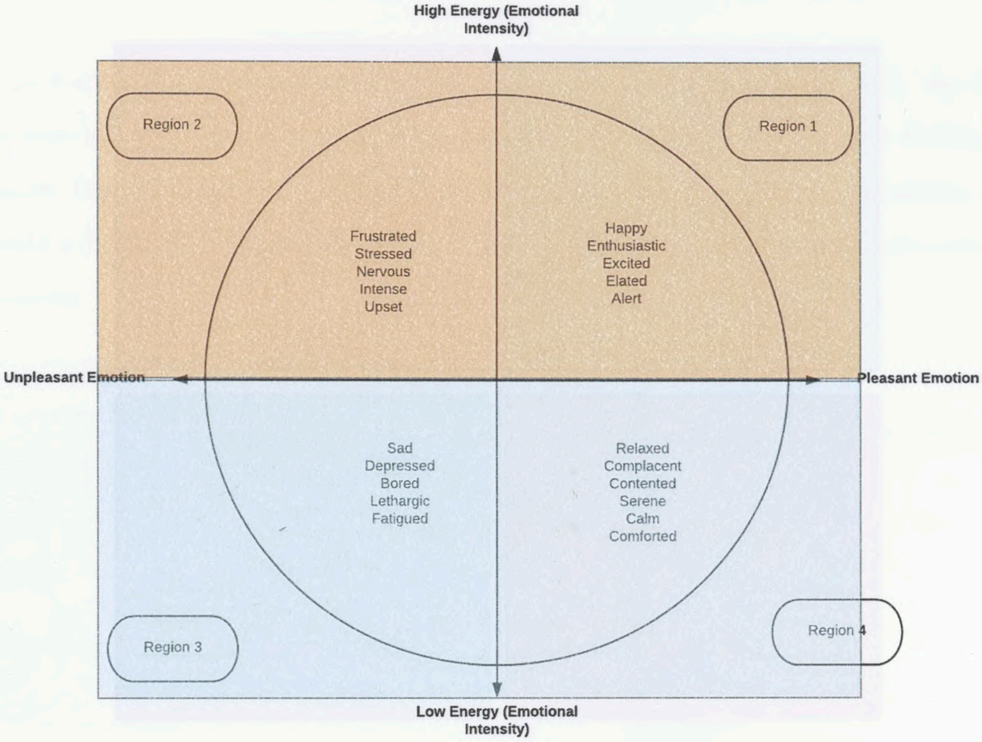


FIGURE 5.10: Valence-level classification accuracy at three comparative training to testing ratios for SSG Burch and SFC Johnson Chatbots. This dimension represents the classification accuracy among all the two regions regions on the Circumplex Model of Affect (i.e., Percentage of utterances correctly classified as either high valence (i.e., Region 1 or Region 4 [unpleasant emotions]) or low valence (i.e., Region 2 or Region 3 [pleasant emotion])



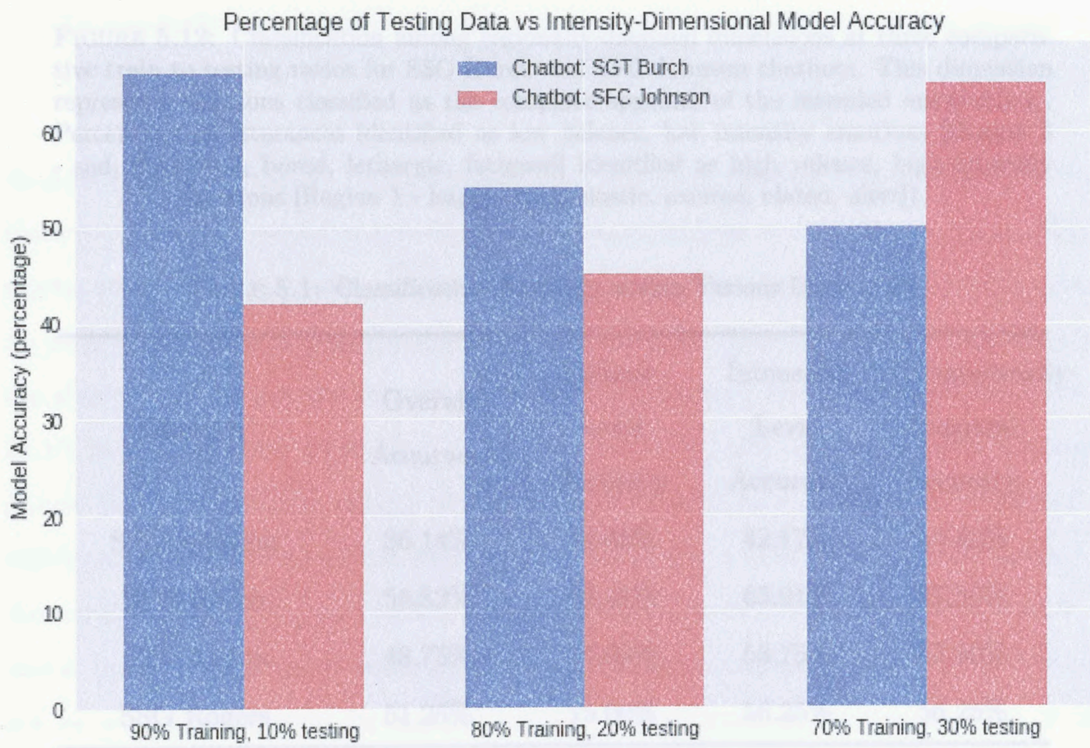


FIGURE 5.11: Classification within the arousal (intensity) dimension at three comparative train to testing ratios for SSG Burch and SFC Johnson Chatbots. This dimension represents the classification accuracy among two regions on the Circumplex Model of Affect (i.e., Percentage of utterances correctly classified as either high intensity (i.e., Region 1 or Region 2 [more highly arousing emotions]) or low intensity (i.e., Region 3 or Region 4 [minimally arousing emotions])

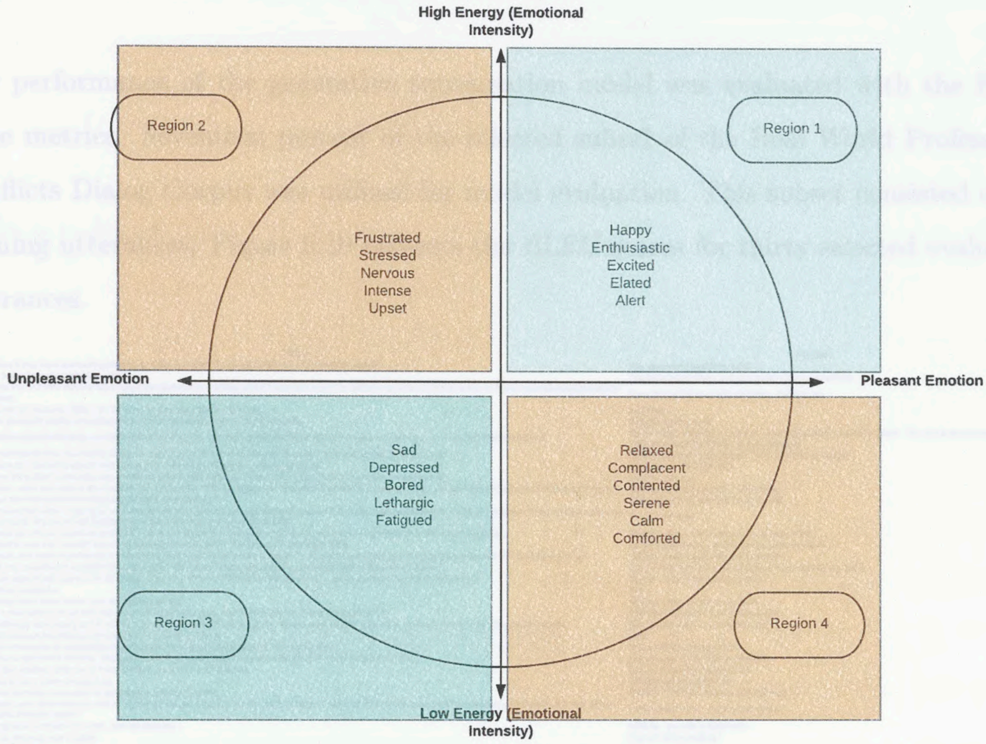


FIGURE 5.12: Classification among regionally-diagonal dimensions at three comparative train to testing ratios for SSG Burch and SFC Johnson chatbots. This dimension represents emotions classified as the complete opposite of the intended emotion(e.g., Percentage of utterances identified as low valence, low intensity emotions [Region 3 - sad, depressed, bored, lethargic, fatigued] identified as high valence, high intensity emotions [Region 1 - happy, enthusiastic, excited, elated, alert]).

TABLE 5.1: Classification Accuracy within Various Dimensions

Chatbot	Overall Accuracy	Valence-Level Accuracy	Intensity-Level Accuracy	Dimensionally-Diagonal Accuracy
SFC Johnson	36.14%	74.69%	42.17%	55.42%
SSG Burch	56.82%	78.41%	65.91%	62.50%
PFC Lewis	48.75%	71.25%	53.75%	57.50%
SSG Rogers	51.25%	75.00%	56.25%	56.25%

This table presents the classification accuracy percentages within four dimensions on the Circumplex Model of Affect; As shown here the highest percentages of correctly classified utterances were within the valence dimension. *These classification results correspond to a 90 percent training to 10 percent testing ratio.

5.5.2 Generative Conversation Model

The performance of the generative conversation model was evaluated with the BLEU score metrics. Seventeen percent of the selected subset of the Real World Professional Conflicts Dialog Corpus was utilized for model evaluation. This subset consisted of 150 training utterances. Figure 5.10 presents the BLEU scores for thirty selected evaluation utterances.

	User	Chatbot	Bleu Score
0	I don't care if you have personal problems at home, you can't fly off the handle on your soldiers, period.	You are a sorry leader yourself ..	0.117084
1	Your no better than the person you train.	You are a sorry leader with the morals of your leadership !.	0.626284
2	This is not the first time that we have heard you lash out at a soldier and if it happens again we will have to take corrective action to find a more permanent solution to the issue.	If you dont mind your own business . I will be out ..	0.154274
3	I did not ask for excuses SSG, I do not want to see that happening again.	You suck ..	0
4	Your wife probably makes you sleep on the couch because of your horrible attitude.	You are a sorry man ..	0
5	That is unacceptable the way you treated your soldier. You cannot take out your frustrations on them. You will have extra staff duty this weekend.	No you dont know me about that . You are a horrible leader yourself . You are a sorry leader . You are a horrible leader yourself yourself ..	0.562341
6	I don't understand how your family could be mad about that. I dont blame you for getting after the soldiers like that at all.	I dont know how to do my job . I know how to do my job ..	0.242281
7	Stop being an intolerable leader and taking your frustration out on your squad members.	It is a direct reflection on it ..	0.331257
8	There one thing that I know you can control and that's your attitude. If I see this again you will be receiving a counseling.	I will not 15 you ..	0.0782156
9	SSG Burch, what you said to the soldiers was deplorable. I will be letting our Company Commander know what you said and have him/her deal with the situation.	you are a sorry example of leadership ..	0.0929544
10	Have you ever thought that it might just be time for a divorce?	Well , you dont know how to do my job ..	0.540854
11	It does not matter what is going on at home. When you took this job, you devoted yourself fully to this job. Just because things get stressful at home does not make this acceptable. If you can not handle the stress of a family, you can't handle this.	You are a retard !.	0.000420253
12	I don't care about your personal problems! You are a soldier and a leader, and need to act more of a professional around your squad.	Wait a minute . Dont tell me how to do my job ..	0.280217
13	If you want to use the "r-word," you must not want to be a leader. You're suspended indefinitely.	I am a professional ! You are a sorry excuse to be !.	0.264543
14	SSG Burch, you need to leave the issues you're having with your family at home. You need to apologize to Jones and go somewhere for a while to cool off.	You are a sorry example of leadership ..	0.0665799
15	If I catch you speaking to a soldier like that again I will have to counsel you and recommend further disciplinary action.	You cannot ! you are a sorry excuse for a treat you to fight ..	0.485987
16	I think you should just leave work for today... be in my office at 0700 in the morning in your dress uniform. You have done it now.	Yes sir . I will . I am a professional ..	0.153935
17	Your wife is ungrateful.	Yes ..	0
18	what a nag! don't listen to her, you did nothing wrong!	Hold up , stop calling me in here ..	0
19	You are a complete fuck up and you just won't stop fucking up. How would you like the same treatment?	Well you dont know how to do my job ..	0.242281
20	You're behavior was wrong and it will be reported today. I want you to go out there and apologize. You will be working overtime for the next few nights.	You suck leader . I am a grown man ..	0.127129
21	That is no excuse for treating him that way. I am writing you up.	Yes !.	0
22	That sort of behavior is absolutely unacceptable. Your failing marriage isn't Jones' problem!	I am a failure of leadership ..	0.318268
23	Why did you yell at him like that. You can't vent off your frustrations at your subordinates. Work on yourself first.	you are a sorry for your information ..	0.265851
24	None of that matters. Get out of here and treat your men with respect!	I refuse to let you handle this !.	0
25	You need to calm down and let it go.	I dont need you telling me ..	0.524736
26	You will be formally counseled for your innapropriate behavior.	I DONT ! ! I DONT have to listen to you ..	0.52684
27	SSG Burch, you were wrong to call him out like that. As a leader you should know better than that. How would you like it if I started to yell at you! I am counseling you for that wrongful action!	I am a grown man . I will not apologize ..	0.0404592
28	dont bring youre problems to work. Just do your job	Hold up , you are a sorry man ..	0
29	Keep it up and your ass is gone.	You are doing nothing !.	0
30	Average Bleu Score	nan	0.202086

FIGURE 5.13: Sample utterances and their corresponding BLEU scores; The average score is 0.20

As previously mentioned, Turing Tests offer a robust measure of chatbot performance. However, they may require a substantial investment of time and resources. Thus, BLEU scores were utilized for the evaluation of the generative chatbot model performance. As seen in Figure 5.10, the average BLEU score for emotional utterances produced by the chatbot within the low valence, high arousal dimension was 0.20. Generally, higher BLEU scores are more desirable (i.e., greater than 0.50) [88]. However, for preliminary phases of model training, with a relatively small dataset, this chatbot's performance is highly promising. Furthermore, while BLEU scores offer a feasible automated metric for conversational modeling evaluation, these scores are generally used for the evaluation of neural machine translation tasks (e.g., translation from English to German) and would not be a sufficient option for chatbot testing. An informal qualitative analysis of the chatbot performance, however, reveals the effectiveness of two-phase model training. This may be extended to a three-phase model training methodology in future extensions of research.

Chapter 6

Conclusions

6.1 Implications and Discussion

This Master's Thesis describes the research aim examined as part of the author's work as a Consortium Research Fellow at the Army Research Institute facilities. Within this work two major conversational modeling tasks are presented: an emotion classification task through the application of Dialog Flow and a Generative Conversational Model through the use of a Neural Machine Translation Model. The classification model predicts the combined emotional valence and emotional intensity of a given utterance upon the Circumplex Model of Affect in a discrete manner. The generative model predicts expressions that are generated in response to a given utterance within regions on the Circumplex Model of Affect, in a regressional manner [4].

Emotional Intelligence is the ability to operate from a place of another's perspective when assessing a given situation. The ability to carry out emotionally intelligent conversation is a highly crucial life skill. For Army leaders, the ability to carry out effective, emotionally intelligent dialogue can be the difference between toxic leadership and effective leadership. The proposed model may offer a way to objectively assess an Army leader's interpersonal strengths and weaknesses in future research endeavors at Army Research Institute Facilities, at the Fort Benning Installation [2]. Future scenario-based assessment tools may leverage generative conversational models in order to facilitate affective dialogue between a human user and chatbot. Furthermore such a model may be used to keep track of long-term dependencies during a scenario-specific conversation, thus allowing for a historical recollection of conversational cues.

The current research project sought to answer three major fundamental questions including: How can a conversation between a virtual agent (character) and human user be modeled? In what way should data be collected in order to generate realistic affective conversational flow? How much training data is needed in order to produce a robust model performance? A couple of informal examined questions include what role do factors like situational complexity and scenario verbosity play in the quality of the overall dataset, and thus model performance [5]?

Examination of the current research aim indicates that conversation can be effectively modeled between a human user and conversational agent through the use of deep learning methodologies. Specifically, Seq2Seq architecture provides an effective way to leverage fluent and contextually appropriate conversation. This architecture will be particularly effective with a large, high-quality dataset [1].

Preliminary results indicate that neural machine translation methodologies may be effectively applied to current and future research projects at ARI. These research projects see to assess interpersonal attributes of leadership in an objective manner. The outcomes of the current thesis research project provide foundation guidelines for current and future research endeavors within this domain.

6.2 Future Directions

Outcomes of the current study were limited due to the use of small, heterogeneous datasets that were not necessarily conversational in nature. Currently openly available emotionally-annotated dialogue corpuses are not as readily available for research purposes as are general conversational corpuses (e.g., OpenSubtitles Dataset [89], Cornell Movie Dataset [19]). Future research endeavors should consider the creation of a larger more high-quality dataset through the use of impromptu-style actors who provide scenario-specific emotional responses. Current researchers at ARI facilities should leverage the work of past ARI investigators who examined the development of interactive interpersonal training tools [84].

These tools were developed through the use of an impromptu-acting style data collection process and provide a good example of how to effectively expand the current dataset in a way that can effectively train a generative conversation model that can model multi-turn fully expressive human to human conversation. The model developed using DialogFlow is limited due to the use of a classification based training methodology, that classifies the user responses and provide a single-turn response to a character in a given scenario based upon that classification. In this way, the model appears to generate erratic responses that are not realistic to everyday emotion expression. The everyday human emotional expression can take several turns within the conversation before invoking shifts in the expressed emotional dynamics. Creating a model on this level may require a substantial investment in time, resources, expertise, and participation among the target users (junior officers in MCoE training facilities).

Furthermore, due to the limited scope of the current research aim a small conversational dialogue corpus was utilized (i.e., Cornell Movie Dialogue Corpus) this dataset could be expanded in order to allow for more robust model performance in future iterations of model training. One example of a high-quality well-formatted dataset that could be used in future research aims is the Open Subtitles Dataset. The Open Subtitles Dataset is one of the largest and most popular corpuses [90]. Particularly useful is the preprocessed version which contains 11.3 million human utterances (each utterance contains at least six words) [78]. This dataset may be leveraged in future phases of the current research aim in order to provide a more robust and less erratic conversational model.

6.2.1 Additional Datasets To Be Used in Future Research

6.2.1.1 ISEAR

The ISEAR database was compiled during the early 1990s and was the result of a project on which psychologist from all over the world collaborated. This project was laid by Klaus R. Scherer and Harold Wallbott. Respondents were asked to report situations in which particular emotions were experienced. These emotions include joy, fear, anger, sadness, disgust, shame, and guilt. Participants appraised the given situation and then provided their verbal reactions to those situations. The final sample was collected from 3000 respondents from 37 countries. Although the current project analysis emotions within four major categories on the Circumplex Model of Affect (i.e., happy/enthusiastic, angry/intense/frustrated, sad/despondent/bored, calm/relaxed); the ISEAR dataset will still be useful. Particularly, when can utilized joy, anger, shame, and guilt as categorical training utterances for the high valence/high intensity, low valence/high intensity, low valence, and the low-intensity revisions of the Circumplex Model of Affect. Hence the only region that this dataset may not adequately address is the high valence, low-intensity region (e.g., representing emotions like calm/relaxed). This weakness will be addressed through the use of multiple datasets that represents varying categorical utterances of emotion [91].

6.2.2 The Valence and Arousal Facebook Posts Dataset

Pietro [92] delivered a dataset of emotional expressions utilizing 3,120 Facebook posts. These Facebook posts were qualitatively examined by human-annotators along the two dimensions of the Circumplex Model of Affect. An Interval scale was utilized for determining the valence (1 [very negative]-5 [neutral]- 9[very positive]) and arousal (1 [neutral, objective] - 9 [very high intensity]) of a given utterance. For the purpose of the current study responses with natural valence scores(i.e., V = 5) may be omitted. Since the current research aims utilize the Circumplex Model of Affect as the foundational framework this dataset may be particularly pertinent to the current research aim. [92]

6.2.3 Clean Balanced Emotional Tweets (CBET) Dataset

The Clean Balanced Emotional Tweets (CBET) dataset contains 81,000 utterances, each labeled with up to two emotions [93]. Emotional labels for this dataset include anger surprise, joy, sadness, fear, disgust, guilt, and thankfulness. This dataset can be used to train utterances within three quadrants on the Circumplex Model of Affect. Quadrant 1, which contains emotions that are higher in valence and higher in intensity (e.g., happy, enthusiastic) may utilize instances with the emotional labels: joy, thankfulness, and love. Quadrant 2, which contains emotions that are lower in valence and higher intensity (e.g., angry, frustrated) may utilize instances with emotional labels: anger, fear, and disgust. Quadrant 3, which contains emotions that are lower in valence and lower in intensity may utilize instances with emotional labels: sadness and guilt [93] [94].

As previously mentioned CBET instances can only be leveraged to train utterances within three quadrants on the Circumplex Model of Affect, the resulting overall dataset may be imbalanced. In order to address this weakness, the Valence and Arousal Facebook Posts Dataset may be used. Particularly, utterances with valence scores between 5 and 9 and intensity scores between 1 and 5 correspond to the Quadrant 4 Emotional Region [93] [94].

6.2.4 Overall

The current research aim explored the use of a generalized neural machine translation based conversational architecture to the modeling of empathy-based conversation. While the current implementation demonstrated the promising potential of this approach further examination should be considered. Namely, the current modeling approach should be extended with the use of pre-trained word2vec word embedding architecture (i.e., Google News Word2Vec)[95]. The use of such a model may slow down model training but will make the overall model performance more robust. As mentioned in chapter 4, word embeddings allow syntactical and semantic relationships to be mapped out during the input pre-processing steps. Overall, this approach may increase the accuracy of the model during training and may yield a more human-to-human level of conversation. However, it should be noted that a generic word2Vec word embedding architecture may have a limited application to the modeling of affective discourse. Thus, it will be important to fine-tune word embeddings to a specific high-quality emotion based corpus during upcoming phases of research.

Further investigation is required in order to increase the robustness of model training. Future research may focus on the creation of a more well-rounded emotion-based dialogue corpus with a real human to the human multi-turn conversation. While DialogFlow facilitated an illuminating analysis for the current research scope, future research may focus on implementing custom made models in python with through the use of TensorFlow Libraries. In the future, model training should be broken into three phases: General Conversation Generation, Emotion-Specific Conversation Generation, and Domain-based(i.e., Scenario) Emotion Specific Conversation, remembering that narrowing the domain of conversation, increases model performance. Overall, results were promising and indicated the effectiveness of current methodologies in supporting empathetic dialogue exchanges within professional Army leadership contexts.

Appendix A

Real-World Professional Conflict Assessment

A.1 Assessment Preview

The following is an example of an instructional scenario included within the assessment that supported the creation of the Real World Professional Conflicts Corpus. The entirety of this assessment can be found within the following GitHub repository: https://github.com/adowdell18/RealWorldProfessionalConflictsDialogCorpus

A.2 Instructions and Exercise Overview

The activity that you will be completing will involve the following tasks: You will be presented with a scenario that will describe the character's emotional state at the beginning of a given scenario. You will then be asked to generate a response that may alter the character's internal emotional state from one position on the Emotion Grid (see Figure 1), to another position (e.g., Region 1, Region 2, Region 3, and Region 4).

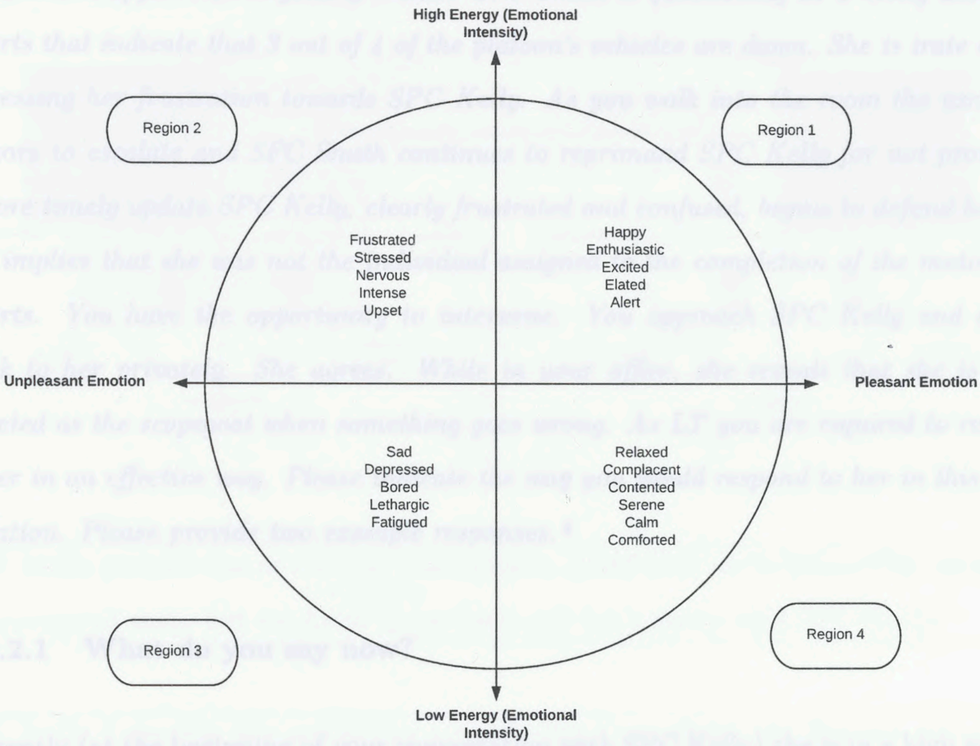


FIGURE A.1: You will be asked to transition the presented characters to emotions listed within the four regions on this grid.

A.2.1 Task

You will be presented with a series of 4 scenarios. After reading each scenario you must generate a unique response that relates to an emotion listed in the specified region of the emotional grid (below).

A.2.2 Example

Instructions: Please read the scenario below and generate several responses, as specified.

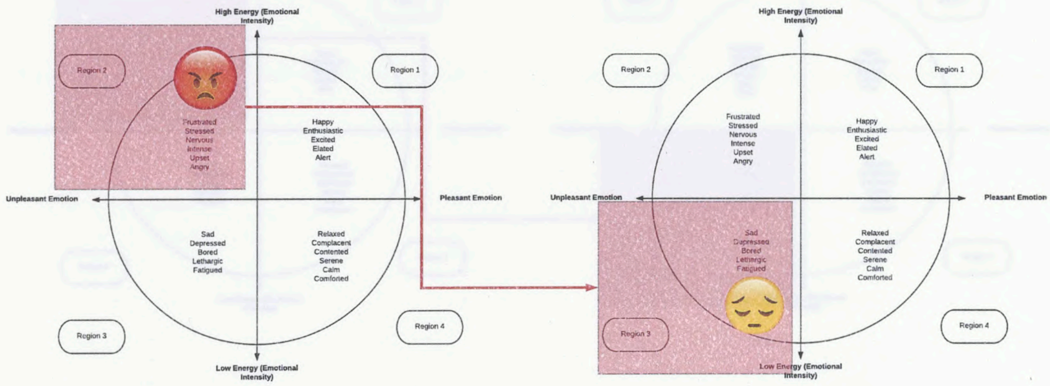
"You have now been in the position as the Platoon Leader of the 1st platoon for one month. You are sitting in your office on Monday afternoon. The entire Company is scheduled to deploy for a field training exercise (FTX) tomorrow morning. The unit is conducting its final checks and preparations. You overhear a conversation directly outside your office between your platoon sergeant, SFC Smith and SPC Kelly, the maintenance

clerk, and it appears to be getting louder. SFC Smith is questioning SPC Kelly about the reports that indicate that 3 out of 4 of the platoon's vehicles are down. She is irate and is expressing her frustration towards SPC Kelly. As you walk into the room the exchange appears to escalate and SFC Smith continues to reprimand SPC Kelly for not providing a more timely update.SPC Kelly, clearly frustrated and confused, begins to defend herself. She implies that she was not the individual assigned to the completion of the motor pool reports. You have the opportunity to intervene. You approach SPC Kelly and ask to speak to her privately. She agrees. While in your office, she reveals that she is often targeted as the scapegoat when something goes wrong. As LT you are required to respond to her in an effective way. Please indicate the way you would respond to her in this given situation. Please provide two example responses."

A.2.2.1 What do you say now?

Currently (at the beginning of your conversation with SPC Kelly) she is in a high energy, unpleasant emotional state: (see region 2 in the image below) frustrated and stressed. What would you say to shift her emotions to the following states? Please provide two responses for each of the four states.

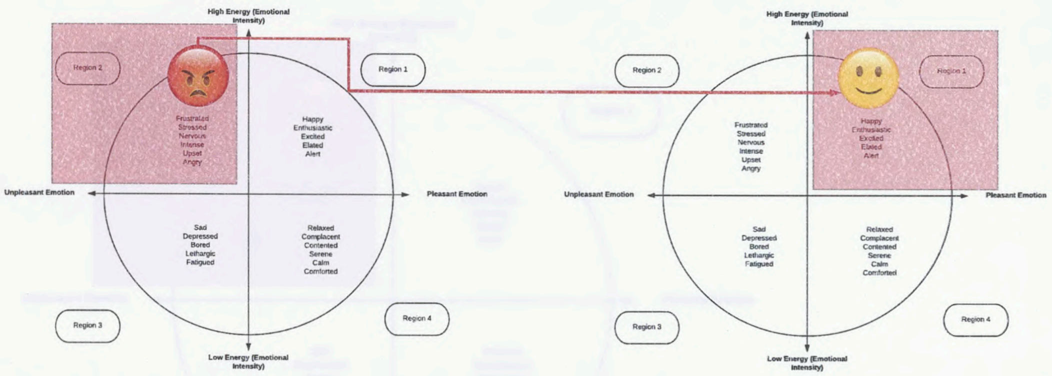
Transition state to low energy, pleasant state



I know that SFC Smith is frustrated but this is not your fault. This is an issue with training. We will all talk about ways we can improve communication within this platoon.

SPC Kelly don't worry too much. Things are going to get better around here soon. Will be working on the systematic communication within this unit.

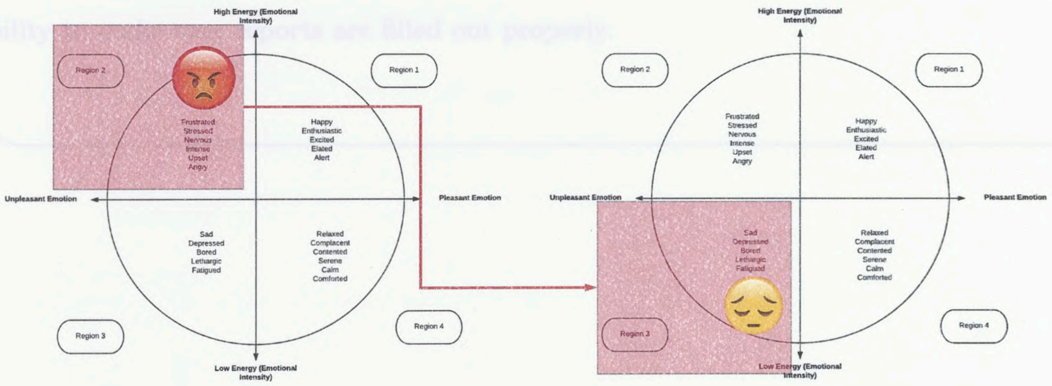
Transition state to high energy, pleasant state



SFC Kelly, you are one of the best maintenance clerks that we've had in this unity. You'll be promoted soon. Let me know if you have any more issues with SFC Smith.

SFC Kelly, I see that you are having a tough day. I am giving you the rest of the day off.

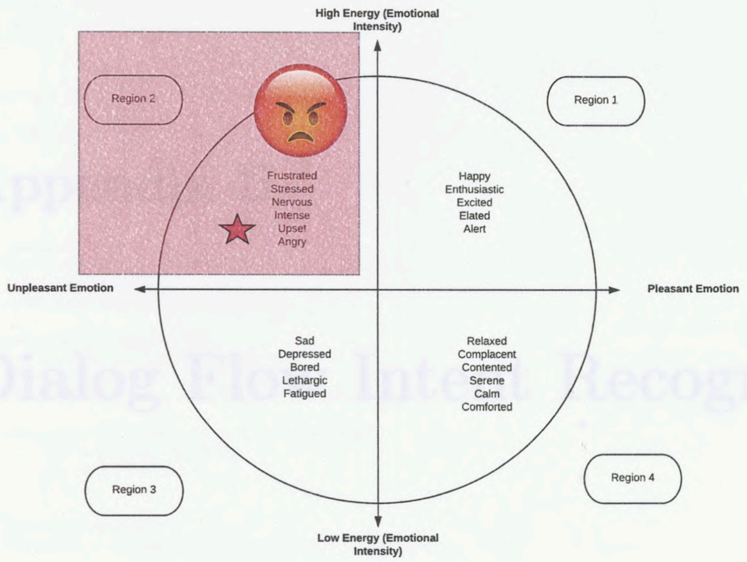
Transition state to low energy, unpleasant state



SFC Kelly, if you want to do well you need to learn to shut your mouth and do what you're told. You are on your last strike right now.

SFC Kelly, get it together. How many times do we have to ask you to follow through with your responsibilities? You are holding our whole unit up with your inconsistent performance.

Maintain high energy, unpleasant state



SFC Kelly, you say that you are the scapegoat but in actuality you work quality is unacceptable.

SFC Kelly, who do you think you are speaking to a superior this way? It is your responsibility to make sure reports are filled out properly.

Appendix B

Dialog Flow Intent Recognition API

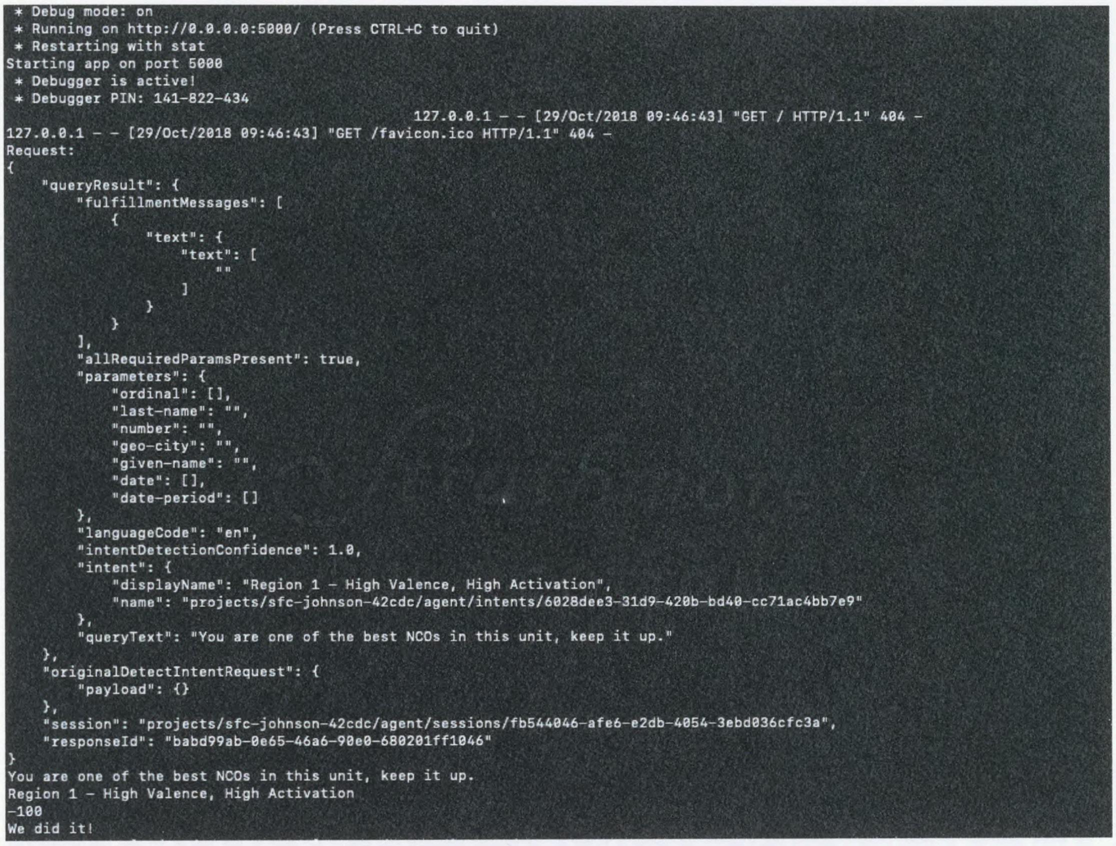
B.1 Description

The current study required the development of an API tool that was used for testing the intent of evaluation utterances. This automated testing tool allowed for speedy model evaluation. An NGROK tool was utilized for interfacing a local server with the DialogFlow server, along with required API keys. The script is detailed below.

```python
import json
import os
from flask import Flask
from flask import request
from flask import make_response
from pprint import pprint
import csv


#flask app should start in global layout

app = Flask(__name__)

#@app.route('/webhook', methods = ['POST'])
@app.route('/butterfly', methods = ['GET', 'POST'])


def webhook():
    req = request.get_json(silent=True, force=True)
    print("Request: ")
    print(json.dumps(req, indent=4))

    print(getUtterance(req))
    print(getActualClassfication(req))
    print(getExpectedClassificaiton())
    ### Allows us to write data to file and then use it

    with open ('dataJSON.js', 'w') as outfile:
        json.dump(req, outfile)
        print("We did it!")
    return r

def getUtterance(req):
    utterance = req.get("queryResult").get("queryText")

    return utterance
def getActualClassfication(req):
    ac = req.get("queryResult").get("intent").get("displayName")
    return ac
def getExpectedClassificaiton():
    ec = -100
    for row in csv_f:
```

```python
    print(getExpectedClassificaiton())
    ### Allows us to write data to file and then use it

    with open ('dataJSON.js', 'w') as outfile:
        json.dump(req, outfile)
        print("We did it!")
    return r

def getUtterance(req):
    utterance = req.get("queryResult").get("queryText")

    return utterance
def getActualClassfication(req):
    ac = req.get("queryResult").get("intent").get("displayName")
    return ac
def getExpectedClassificaiton():
    ec = -100
    for row in csv_f:
        if row[0] == getUtterance:
            ec = row[1]
    return ec

def printUserInput(req):
    #response = str(req.get("result").get("resolvedQuery"))
    return response

if __name__ == '__main__':
        #string = "Hey!"
        #print(string)
        writeLine = []
        f = open('test-90.csv')
        csv_f = csv.reader(f)
        port = int(os.getenv("PORT", 5000))
        print("Starting app on port %d" %(port))
        app.run(debug = True, port = port, host = '0.0.0.0')

        print("Starting app on port %d" % (port))
```

FIGURE B.1: Python script that allowed a local server to pass text based utterances within an inputted CSV to a DialogFlow server; script outputs a CSV of corresponding intents



```
 * Debug mode: on
 * Running on http://0.0.0.0:5000/ (Press CTRL+C to quit)
 * Restarting with stat
Starting app on port 5000
 * Debugger is active!
 * Debugger PIN: 141-822-434
                                                127.0.0.1 - - [29/Oct/2018 09:46:43] "GET / HTTP/1.1" 404 -
127.0.0.1 - - [29/Oct/2018 09:46:43] "GET /favicon.ico HTTP/1.1" 404 -
Request:
{
    "queryResult": {
        "fulfillmentMessages": [
            {
                "text": {
                    "text": [
                        ""
                    ]
                }
            }
        ],
        "allRequiredParamsPresent": true,
        "parameters": {
            "ordinal": [],
            "last-name": "",
            "number": "",
            "geo-city": "",
            "given-name": "",
            "date": [],
            "date-period": []
        },
        "languageCode": "en",
        "intentDetectionConfidence": 1.0,
        "intent": {
            "displayName": "Region 1 - High Valence, High Activation",
            "name": "projects/sfc-johnson-42cdc/agent/intents/6028dee3-31d9-420b-bd40-cc71ac4bb7e9"
        },
        "queryText": "You are one of the best NCOs in this unit, keep it up."
    },
    "originalDetectIntentRequest": {
        "payload": {}
    },
    "session": "projects/sfc-johnson-42cdc/agent/sessions/fb544046-afe6-e2db-4054-3ebd036cfc3a",
    "responseId": "babd99ab-0e65-46a6-90e0-680201ff1046"
}
You are one of the best NCOs in this unit, keep it up.
Region 1 - High Valence, High Activation
-100
We did it!
```

FIGURE B.2: Terminal view of DialogFlow python/flask interface via Ngrok web services.

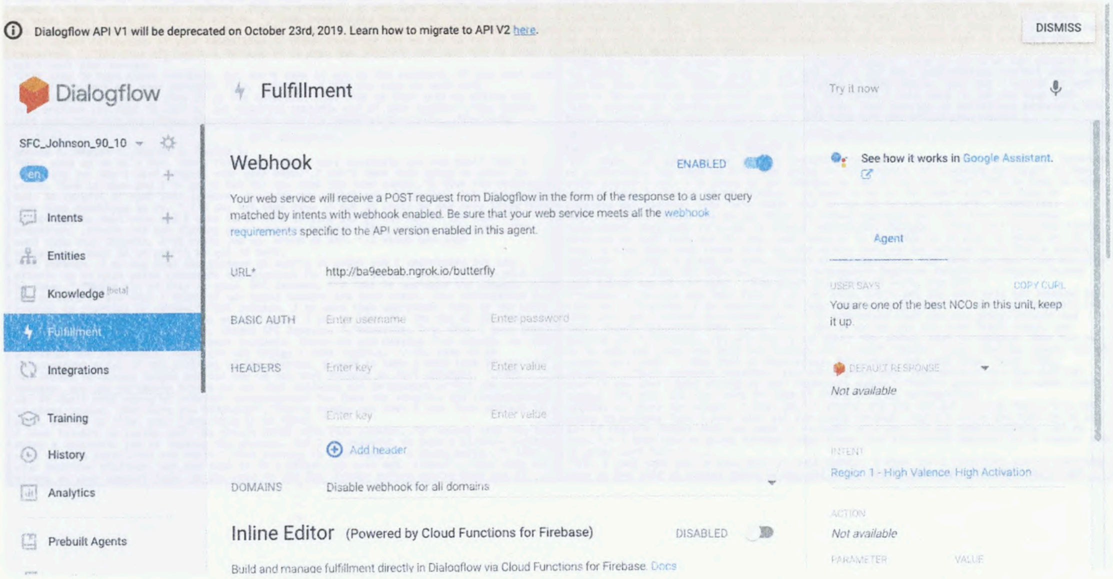
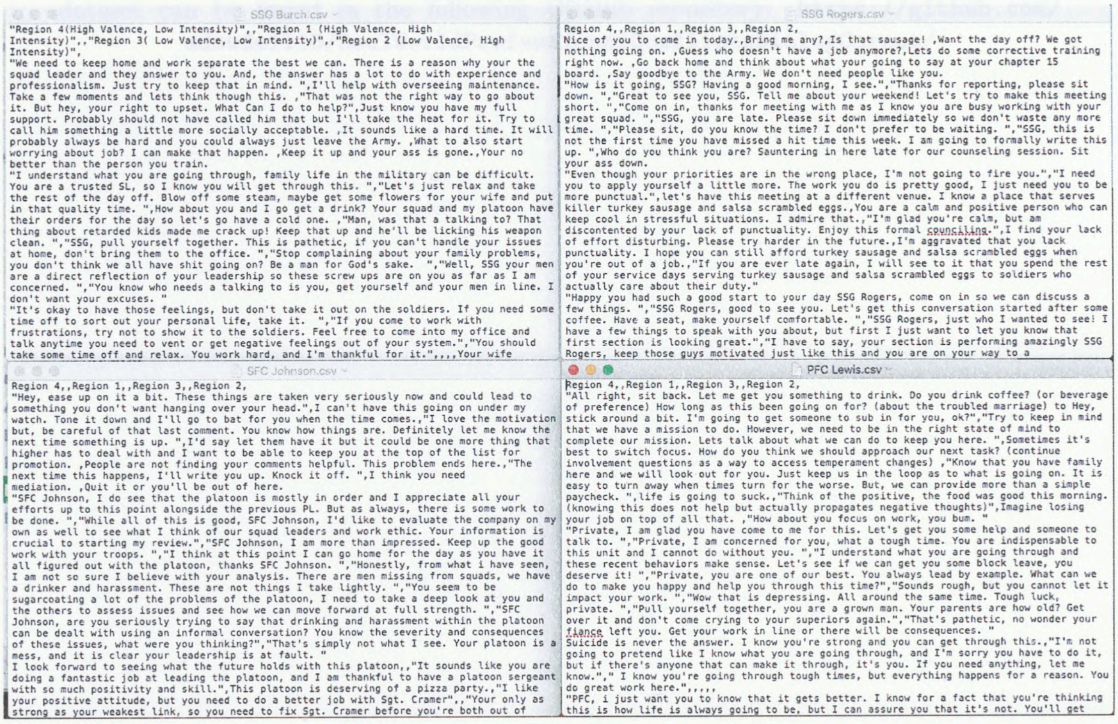


FIGURE B.3: Web interface view of Dialog Flow web-hook fulfillment. This fulfillment allows for the transmission of JSON data via a server.

Appendix C

Real World Professional Conflicts Dialog Corpus



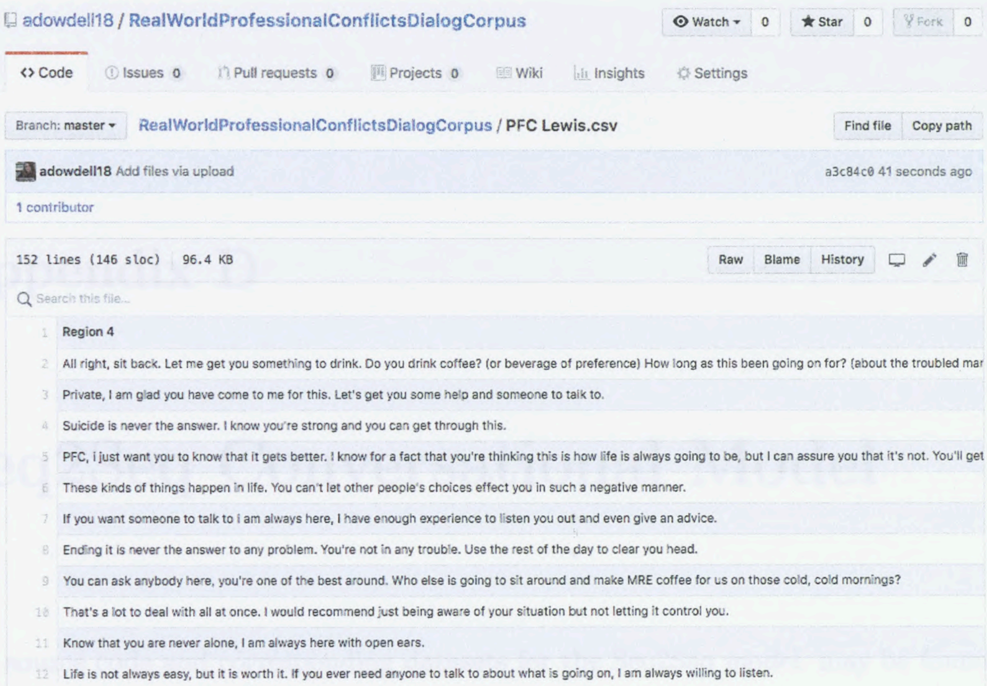


FIGURE C.1: A corpus of emotional utterances were compiled for this project. The dataset can be found in the following GitHub repository: `https://github.com/adowdell18/RealWorldProfessionalConflictsDialogCorpus/`

Appendix D

Seq2Seq Conversational Model

The source code and corresponding datasets for the Seq2Seq model, may be found here `https://github.com/adowdell18/Seq2Seq-Conversation-Models`. Code and comments were provided, sourced, and/or modified from the following Udemy-tutorial inspired repositories: `https://github.com/AbrahamSanders/seq2seq-chatbot`, `https://github.com/lucko515/chatbot-startkit`, `https://www.udemy.com/chatbot/`.

```
        epoch_accuracy.append(get_accuracy(np.array(y_batch), np.array(preds)))
        bucket_accuracy.append(get_accuracy(np.array(y_batch), np.array(preds)))

        bucket_loss.append(cost)
        epoch_loss.append(cost)
        #for s in preds:
          #print("Chatbot: ",int2str(s))
    #saver.save(session, "gdrive/My Drive/checkpoint/epoch{}/chatbot_{}.ckpt".format(i,b))
    print("Bucket {}:".format(b+1),
          " | Loss: {}".format(np.mean(bucket_loss)),
          " | Accuracy: {}".format(np.mean(bucket_accuracy)))
BucketID=0
```

```
100%|██████████| 2/2 [00:04<00:00,  2.54s/it]
  0%|          | 0/2 [00:00<?, ?it/s]
Bucket 1:  | Loss: 2.521085023880005  | Accuracy: 0.134375
100%|██████████| 2/2 [00:01<00:00,  1.23it/s]
  0%|          | 0/2 [00:00<?, ?it/s]
Bucket 1:  | Loss: 2.472428798675537  | Accuracy: 0.14010416666666667
100%|██████████| 2/2 [00:01<00:00,  1.22it/s]
  0%|          | 0/2 [00:00<?, ?it/s]
Bucket 1:  | Loss: 2.3946547508239746  | Accuracy: 0.14375
100%|██████████| 2/2 [00:01<00:00,  1.23it/s]
  0%|          | 0/2 [00:00<?, ?it/s]
Bucket 1:  | Loss: 2.3234572410583496  | Accuracy: 0.13932291666666666
100%|██████████| 2/2 [00:01<00:00,  1.22it/s]
  0%|          | 0/2 [00:00<?, ?it/s]
Bucket 1:  | Loss: 2.242194414138794  | Accuracy: 0.14036458333333335
100%|██████████| 2/2 [00:01<00:00,  1.23it/s]
  0%|          | 0/2 [00:00<?, ?it/s]
Bucket 1:  | Loss: 2.171445369720459  | Accuracy: 0.14244791666666667
100%|██████████| 2/2 [00:01<00:00,  1.23it/s]
  0%|          | 0/2 [00:00<?, ?it/s]
```

```
#The following code and comments are provided, sourced, and/or modified from the following repositories from: https://github.com/AbrahamSanders/seq2seq-ch
for i in range(0,EPOCHS):
    epoch_accuracy = []
    epoch_loss = []
    for b in range(0,len(bucketed_data)):
        bucket = bucketed_data[b]
        questions_bucket = []
        answers_bucket = []
        bucket_accuracy = []
        bucket_loss = []
        for k in range(len(bucket)):
            questions_bucket.append(np.array(bucket[k][0]))
            answers_bucket.append(np.array(bucket[k][1]))
        #for ii in tqdm(range(len(questions_bucket) //  BATCH_SIZE)):
        Number_of_Loop = len(questions_bucket) / BATCH_SIZE
        Number_of_Loop = math.trunc(Number_of_Loop)
        for ii in tqdm(range(Number_of_Loop)):

        #for ii in tqdm(range(len(questions_bucket) //  BATCH_SIZE)):

            starting_id = ii * BATCH_SIZE

            X_batch = questions_bucket[starting_id:starting_id+BATCH_SIZE]
            y_batch = answers_bucket[starting_id:starting_id+BATCH_SIZE]

            feed_dict = {model.inputs:X_batch,
                         model.targets:y_batch,
                         model.keep_probs:KEEP_PROBS,
                         model.decoder_seq_len:[len(y_batch[0])]*BATCH_SIZE,
                         model.encoder_seq_len:[len(X_batch[0])]*BATCH_SIZE}

            cost, _, preds = session.run([model.loss, model.opt, model.predictions], feed_dict=feed_dict)

            epoch_accuracy.append(get_accuracy(np.array(y_batch), np.array(preds)))
            bucket_accuracy.append(get_accuracy(np.array(y_batch), np.array(preds)))

            bucket_loss.append(cost)
            epoch_loss.append(cost)
            #for s in preds:
              #print("Chatbot: ",int2str(s))
        #saver.save(session, "gdrive/My Drive/checkpoint/epoch{}/chatbot_{}.ckpt".format(i,b))
        print("Bucket {}:".format(b+1),
              " | Loss: {}".format(np.mean(bucket_loss)),
              " | Accuracy: {}".format(np.mean(bucket_accuracy)))
    BucketID=0
```

FIGURE D.1: Code Snippets from Seq2Seq Conversation Model [1]

Bibliography

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[2] US Army. Army doctrine publication 6-22, army leadership, 2012.

[3] Daniel L Schwartz and Dylan Arena. *Measuring what matters most: Choice-based assessments for the digital age*. MIT Press, 2013.

[4] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[5] Randy Brou. Brou.pdf. `https://ndiastorage.blob.core.usgovcloudapi.net/ndia/2018/armyst/Brou.pdf`, . (Accessed on 12/28/2018).

[6] Deep learning and nlp a-z™: How to create a chatbot | udemy. `https://www.udemy.com/chatbot/`. (Accessed on 01/14/2019).

[7] Armin Seyeditabari, Narges Tabari, and Wlodek Zadrozny. Emotion detection in text: a review. *arXiv preprint arXiv:1806.00674*, 2018.

[8] Peter Robinson and Tadas Baltrusaitis. Empirical analysis of continuous affect. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 288–294. IEEE, 2015.

[9] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011.

[10] Charles Darwin and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[11] Jonathan Posner, James A Russell, and Bradley S Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734, 2005.

[12] Ivar FT van Willigen. Reasoning about emotions. 2009.

[13] Michelle Yik, James A Russell, and James H Steiger. A 12-point circumplex structure of core affect. *Emotion*, 11(4):705, 2011.

[14] Mark A Thornton and Diana I Tamir. Mental models accurately predict emotion transitions. *Proceedings of the National Academy of Sciences*, page 201616056, 2017.

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

[16] Clark Elliott. *The Affective Reasoner: A Process Model of Emotions in a Multiagent System.* PhD thesis, PhD thesis, Northwestern University, May 1992. The Institute for the ..., 1989.

[17] Timothy Bickmore and Amanda Gruber. Relational agents in clinical psychiatry. *Harvard review of psychiatry*, 18(2):119–130, 2010.

[18] Hugo Liu, Henry Lieberman, and Ted Selker. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 125–132. ACM, 2003.

[19] Cristian Danescu-Niculescu-Mizil and Lillian Lee. 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 2011*, 2011.

[20] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204, 2010.

[21] Saif M Mohammad and Peter D Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[22] Jacopo Staiano and Marco Guerini. Depechemood: a lexicon for emotion analysis from crowd-annotated news. *arXiv preprint arXiv:1405.1605*, 2014.

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

[24] Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Citeseer, 1999.

[25] Rafael A Calvo and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, 1(1):18–37, 2010.

[26] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. Semeval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, 2018.

[27] Chenyang Huang, Osmar Zaiane, Amine Trabelsi, and Nouha Dziri. Automatic dialogue generation with expressed emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 49–54, 2018.

[28] Ameeta Agrawal. Enriching affect analysis through emotion and sarcasm detection. 2018.

[29] Alexandra Balahur, Jesús M Hermida, Andrés Montoyo, and Rafael Muñoz. Emotinet: A knowledge base for emotion detection in text built on the appraisal theories. In *International Conference on Application of Natural Language to Information Systems*, pages 27–39. Springer, 2011.

[30] Petra Fagerberg, Anna Ståhl, and Kristina Höök. emoto: emotionally engaging interaction. *Personal and Ubiquitous Computing*, 8(5):377–381, 2004.

[31] Kwang In Kim, Keechul Jung, and Jin Hyung Kim. Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1631–1639, 2003.

[32] Weiyuan Li and Hua Xu. Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications*, 41(4):1742–1749, 2014.

[33] Jared Suttles and Nancy Ide. Distant supervision for emotion classification with discrete binary values. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 121–136. Springer, 2013.

[34] Rakesh C Balabantaray, Mudasir Mohammad, and Nibha Sharma. Multi-class twitter emotion classification: A new approach. *International Journal of Applied Information Systems*, 4(1):48–53, 2012.

[35] Maryam Hasan, Elke Rundensteiner, and Emmanuel Agu. Emotex: Detecting emotions in twitter messages. 2014.

[36] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Textual affect sensing for sociable and expressive online communication. In *International Conference on Affective Computing and Intelligent Interaction*, pages 218–229. Springer, 2007.

[37] Erik Tromp and Mykola Pechenizkiy. Rbem: a rule based approach to polarity detection. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 8. ACM, 2013.

[38] CY Yam. Emotion detection and recognition from text using deep learning, 2015. URL `https://www.microsoft.com/developerblog/2015/11/29/emotion-detection-and-recognition-from-text-using-deep-learning/`.

[39] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[40] Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. Semeval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)*, pages 1–18, 2016.

[41] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*, 2017.

[42] Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 174–184, 2018.

[43] WEI Honghao, Yiwei Zhao, and Junjie Ke. Building chatbot with emotions.

[44] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. *arXiv preprint arXiv:1704.01074*, 2017.

[45] Grace Gee and Eugene Wang. psyml at semeval-2018 task 1: Transfer learning for sentiment and emotion analysis. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 369–376, 2018.

[46] Saif M Mohammad and Felipe Bravo-Marquez. Wassa-2017 shared task on emotion intensity. *arXiv preprint arXiv:1708.03700*, 2017.

[47] Sayyed M Zahiri and Jinho D Choi. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. *arXiv preprint arXiv:1708.04299*, 2017.

[48] Muhammad Abdul-Mageed and Lyle Ungar. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 718–728, 2017.

[49] Kiyoaki Shirai, Virach Sornlertlamvanich, Sanparith Marukata, et al. Recurrent neural network with word embedding for complaint classification. In *Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016)*, pages 36–43, 2016.

[50] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. *arXiv preprint arXiv:1811.00405*, 2018.

[51] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 873–883, 2017.

[52] Srini Janarthanam. Hands-on chatbots and conversational ui development: Build chatbots and voice user interfaces with chatfuel, dialogflow, microsoft bot framework, twilio, and alexa skills. 2017.

[53] Nicole M Radziwill and Morgan C Benton. Evaluating quality of chatbots and intelligent conversational agents. *arXiv preprint arXiv:1704.04579*, 2017.

[54] Michael McTear. Conversational modelling for chatbots: Current approaches and future directions. Technical report, Technical report, Ulster University, Ireland, 2018.

[55] Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, et al. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):530–539, 2015.

[56] Nikhil Buduma and Nicholas Locascio. *Fundamentals of deep learning: Designing next-generation machine intelligence algorithms.* " O'Reilly Media, Inc.", 2017.

[57] Kyo-Joong Oh, Dongkun Lee, Byungsoo Ko, and Ho-Jin Choi. A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation. In *Mobile Data Management (MDM), 2017 18th IEEE International Conference on*, pages 371–375. IEEE, 2017.

[58] Shanliang Yang, Qi Sun, Huyong Zhou, and Zhengjie Gong. A multi-layer neural network model integrating bilstm and cnn for chinese sentiment recognition. In *Proceedings of the 2018 International Conference on Computing and Artificial Intelligence*, pages 23–29. ACM, 2018.

[59] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

[60] Sumit Chopra, Michael Auli, and Alexander M Rush. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, 2016.

[61] Li Dong, Furu Wei, Ming Zhou, and Ke Xu. Question answering over freebase with multi-column convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 260–269, 2015.

[62] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. A deeper look into sarcastic tweets using deep convolutional neural networks. *arXiv preprint arXiv:1610.08815*, 2016.

[63] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436, 2015.

[64] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*, 2016.

[65] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

[66] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

[67] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 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*, pages 2539–2544, 2015.

[68] Ozan Irsoy and Claire Cardie. Opinion mining with deep recurrent neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 720–728, 2014.

[69] Duyu Tang, Bing Qin, and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432, 2015.

[70] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 467–474. ACM, 2015.

[71] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[72] Arpan Bhattacharya. Generative conversational agents-the state-of-the-art and the future of intelligent conversational systems. *International Journal on Recent and Innovation Trends in Computing and Communication*, 5(5):817–821, 2017.

[73] Martino Mensio. *Deep Semantic Learning for Conversational Agents.* PhD thesis, Politecnico di Torino, 2018.

[74] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, L Sutskever, and G Zweig. word2vec. *URL https://code. google. com/p/word2vec*, 2013.

[75] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.

[76] Github - farizrahman4u/seq2seq: Sequence to sequence learning with keras. `https://github.com/farizrahman4u/seq2seq`. (Accessed on 01/09/2019).

[77] Yukun Ma, Haiyun Peng, and Erik Cambria. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. In *Proceedings of AAAI*, 2018.

[78] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*, 2016.

[79] Huyen Nguyen, David Morales, and Tessera Chin. A neural chatbot with personality. Technical report, Stanford University working paper, 2017.

[80] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

[81] Angie Dowdell. Real-world professional conflict scenarios: Transitioning emotional states - 7/30/2018, 2018. URL `http://columbusstate.qualtrics.com/jfe/form/SV_42D2uUrZCJ0cxpz`.

[82] Realcareer™ employability skills program work-place scenario cards, 2015. URL `https://www.realityworks.com/ftp_uploads/sample-cards-test.pdf`.

[83] Randy Brou. Building automated assessments of interpersonal leadership skills. `http://www.iitsec.org/-/media/sites/iitsec/agenda/2018-abstract-book.ashx?la=en`, . (Accessed on 12/28/2018).

[84] James E Schroeder, Frederick N Dyer, Paul Czerny, Edward W Youngling, and Daniel P Gillotti. Videodisc interpersonal skills training and assessment (vista). volume 3. scenarios. Technical report, LITTON MELLONICS SYSTEMS DEVELOPMENT DIV FORT BENNING GA, 1986.

[85] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[86] Mitchell Vibbert, Jacques-Olivier Goussard, Richard J Beaufort, and Benjamin P Monnahan. Dialog flow management in hierarchical task dialogs, September 19 2017. US Patent 9,767,794.

[87] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[88] Ying Zhang, Stephan Vogel, and Alex Waibel. Interpreting bleu/nist scores: How much improvement do we need to have a better system? In *LREC*, 2004.

[89] Jörg Tiedemann. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248, 2009.

[90] Pierre Lison and Jörg Tiedemann. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. 2016.

[91] KR Scherer and H Wallbott. International survey on emotion antecedents and reactions (isear)(1990), 2017.

[92] Daniel Preoţiuc-Pietro, H Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. Modelling valence and arousal in facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, 2016.

[93] Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):25, 2017.

[94] Ameneh Gholipour Shahraki and Osmar R Zaiane. Lexical and learning-based emotion mining from text. In *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing*, 2017.

[95] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

I have submitted this thesis in partial fulfillment of the requirements for the degree of Master of Science

1-26-2019

Date

Angie A Dowdell

We approve of the thesis of Angie A Dowdell as presented here.

2/12/19

Date

Rania Hodhod, Ph.D.
Assistant Chair
Associate Professor of
Computer Science, Thesis Advisor

2/26/19

Date

Randy Brou, Ph.D.
Adjunct Professor of Psychology
Research Psychologist
Army Research Institute (ARI)

2/2/19

Date

Shamim Khan, Ph.D.
Professor of Computer Science

2/19/19

Date

Shuangbao Wang, Ph.D.
Professor of Computer Science

2/18/19

Date

Wayne Summers, Ph.D.
Distinguished Chairperson
Professor of Computer Science