When your face and tone of voice don't say it all: Inferring emotional state from word semantics and conversational topics.

Workshop on Cognitive Architectures for Human-Robot Interaction

Andrew Valenti
Tufts University
Human-Robot Interaction Laboratory
Medford, Massachusetts
andrew.valenti@tufts.edu

Meia Chita-Tegmark
Tufts University
Human-Robot Interaction Laboratory
Medford, Massachusetts
mihaela.chita_tegmark@tufts.edu

Theresa Law
Tufts University
Human-Robot Interaction Laboratory
Medford, Massachusetts
theresa.law@tufts.edu

Alexander Bock
Tufts University
Human-Robot Interaction Laboratory
Medford, Massachusetts
alexander.bock@tufts.edu

Bradley Oosterveld
Tufts University
Human-Robot Interaction Laboratory
Medford, Massachusetts
bradley.oosterveld@tufts.edu

Matthias Scheutz
Tufts University
Human-Robot Interaction Laboratory
Medford, Massachusetts
matthias.scheutz@tufts.edu

ABSTRACT

In typical human interactions emotional states are communicated via a variety of modalities such as auditory (through speech), visual (through facial expressions) and kinesthetic (through gestures). However, one or more modalities might be compromised in some situations, as in the case of facial masking in Parkinson's disease (PD). In these cases, we need to focus the communication and detection of emotions on the reliable modalities, by inferring emotions from what is being said, and compensate for the modalities that are problematic, by having another agent (e.g., a robot) provide the missing facial expressions. We describe the initial development stage of a cognitive robotic architecture that can assist the communication and detection of emotions in interactions where some modalities are totally or partially compromised. We hypothesize that the distribution of topics extracted from each sentence, that is part of a collection of written text documents, using the Latent Dirichlet Allocation (LDA) generative model can be associated with measures of emotional valence and arousal. We integrated our model into the cognitive robotic architecture, DIARC, and demonstrated how a robot can use speech transcriptions to detect positive or negative emotion valence and express it through its facial features. This work forms the basis for developing a more robust model with finer prediction resolution. In addition, the model can be further developed to form an auxiliary natural language understanding pipeline which can be used to supplement existing components in the architecture to utilize affect in human-robot interactions.

KEYWORDS

Parkinson's disease; Latent Dirichlet Allocation; affective computing; machine learning; neural networks; topic modeling

1 INTRODUCTION

In typical human interactions, emotions are communicated via a variety of modalities: auditory (through voice intonation and content of speech), visual (through facial expressions and posture), and kinesthetic (through facial and body gestures). In some situations, like in the case of disorders that affect communication, one or more modalities may not be available or reliable for conveying emotions (e.g., visual input from facial expressions). One symptom of Parkinson's disease (PD), a neurodegenerative disorder, is facial masking (hypomimia) which arises from diminished control of one's facial and vocal expression. Facial masking can lead to dissociation between one's inner emotional state and outward facial appearance (e.g., looking angry when one is not). Since several modalities for communicating emotions might be at least partially compromised in PD due to poor vocal and facial control, the content of speech can be used as an alternative for the expression and detection of emotions. For humans who are inherently multi-modal in their detection and expression of emotion, it can be very difficult to override their instincts and ignore certain modalities and thus artificial cognitive architectures could be designed to help with this. Our goal is to create an assistive tool embodied in a robot that can express affect (e.g., by displaying different facial expressions) inferred from verbal content.

We are in the process of developing a system to supplement the natural language understanding pipeline in the DIARC [8] cognitive robotic architecture to automatically detect the emotion dynamics in conversations from what is being said rather than from facial expressions or voice. In our prior research [12], we introduced a novel approach, topic modeling, for automating the detection of emotional content as a conversation unfolds. To express the emotions detected by the topic model, we have created a set of scripted facial expressions used by DIARC to actuate the robot. At present, the model predicts coarse, binary measures of valence or arousal, i.e., positive or negative. Finer degrees of emotion expression need to be captured in the model in order to replicate the rise and fall of emotional state in human social interactions. In future work, DIARC will be able to coordinate evolving affective states with the task driven dialogue system of DIARC to enable more realistic and engaging human-robot interactions.

The paper proceeds as follows. In the *Background and related work* section, we summarize the approaches that have been used so far to evaluate the emotional state of persons with PD. In the *Methods* section we describe how we obtained affect annotations for text sentences drawn from interview transcriptions. The *Architecture*

section reviews how the trained model and associated components were implemented in the DIARC cognitive robotic architecture [8]. We next describe how we demonstrated the emotion pipeline using previously unseen sentences from interview transcriptions and informally evaluated its performance. We then discuss the advantages, disadvantages, and limitations of this approach and the potential for embedding the model in an emotion detecting assistive conversation tool.

2 BACKGROUND AND RELATED WORK

The outward expression of human emotion and feeling is commonly referred to as affect. We can consider representing affect as a point within a two-dimensional circular space in which the x-axis measures valence, emotion positivity, and the y-axis measures arousal, the emotion intensity [3]. This dimensional model of emotion arises from the theory that suggests a common neurophysiological system is responsible for all affective states [7]. Affective computing is a research area investigating automated methods to recognize human affect, expressed, for example, through voice, face, or other biological channels and different modalities [9].

There are models of affect and arousal that traditionally employed sources of information such as physiology, facial expressions, and intonation; for a review of these approaches, see [2] and [1] for one such example where voice and facial expressions were combined with a high success rate. Since the auditory and visual modalities for communicating emotional states are problematic for people with PD due to lack of facial and vocal expressiveness, researchers have proposed that a more accurate way would be to focus on the words a person uses in their verbal or written speech [5]. To date, approaches for detecting affect in persons with PD have commonly used sentiment lexicons e.g., LIWC [6]. These approaches rely on human-generated dictionaries of words and symbols in a particular language that have been shown to correspond to positive and negative emotion categories.

Our prior research [12] suggests that sentiment lexicons can be ineffective when trying to track the emotional content of a conversation as it unfolds. Approaches such as LIWC are not optimal when using as input short pieces of text with few words, which is often the case with speech utterances. To automatically track the progression of the emotional content in a conversation, we have developed a novel method which tries to fit the utterance into the thematic structure of a set of documents on which the model was previously trained [11]. One might think of these documents as representing the conversational domain in which the artificial system can be expected to operate.

We used the Latent Dirichlet Allocation (LDA) generative model as the basis for an unsupervised learning model, which we trained to extract topic proportions from a collection of written text documents. When an unseen sentence was presented to the model, it found its topic proportions and used them as a set of features. We then used a Multi-Layer Perceptron (MLP) classifier to associate these features with training data labeled according to emotion valence (positive or negative) and arousal (high or low). Zhang et al. [13] used a neural network to detect affect from facial expression and a Latent Semantic Analysis model [4], a non-generative predecessor of LDA, to detect topics embedded within the human-robot conversation; however, the detected topics were not used to inform affect detection.

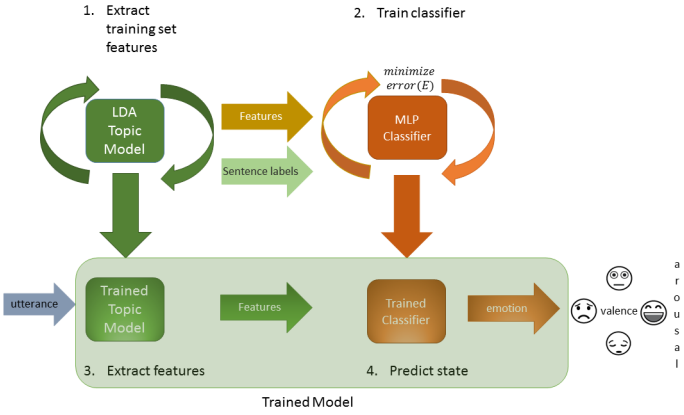

Figure 1: Affective prediction model. (1) The LDA model is trained to extract features from the interview documents (2) For each document's sentence, its topic proportions (features) are extracted and, along with its emotion target, is used to train the classifier (3) Features are extracted from the utterance by the trained LDA model, and (4) presented to the trained classifier to predict its emotional state

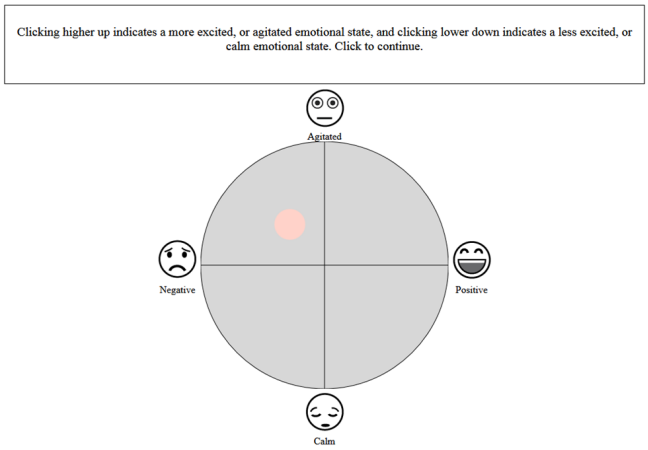

Figure 2: Human-labeling of text using EMotion Inference Tool: EMIT. The mouse is used to position the colored cursor anywhere in the field circumscribed by the emotional circle. At a mouse-click, the tool records the valence & arousal coordinates of the cursor and the elapsed time.

3 METHODS

Our model was trained on the individual sentences drawn from 448 documents with an average word count of 258 words, the largest containing 1732 and the smallest, 2 [12]. The documents were constructed from selected interview transcripts from 106 male and

female participants with PD, living in the community, who participated in a study [10] which asked them to recall two types of experiences: a frustrating one and an enjoyable one that they had during the past week. Thus, the robot could be expected to accurately predict emotion from utterances spoken in a similar contextual domain.

We used the interview documents to collect ratings of the emotional content (i.e., valence and arousal) for each sentence of the document; these values were used as the training targets for the model. We collected this data using Amazon Mechanical Turk (AMT). AMT workers used a Web-based application we created to indicate their perceived emotion contained in text content (see Figure 2). The Web application is described in detail elsewhere [11]. To ensure high-quality of the training data, we only used those sentences for which at least 80% of the raters agreed on the label for valence (positive or negative) and arousal (high, low). For valence this represented 1,058 sentences; for arousal 615 sentences. This shows, as expected, that humans had more difficulty inferring arousal than valence in this dataset.

The model design consists of two processing steps: (i) extract the topic proportions from each document in the set (items 1 and 3 in Figure 1), and (ii) use these features to predict the emotion valence and arousal of individual sentences as yet unseen by the model (items 2 and 4 in Figure 1). Training of the LDA model and the classifier (items 1 and 2 in Figure 1) was done outside of the robotic architecture and the results saved to files. These were subsequently used to initialize the emotion pipeline. In principle, training could also take place in the robotic architecture.

We trained the LDA topic model to generate vectors with 34 features. We used a Multi-layer Perceptron (MLP) with one hidden layer and 34 artificial neurons to associate the features produced by the trained LDA model with the training targets collected from the human workers. Further detail on the model design and parameter selection is given in [11]. MLPs can be configured as multi-label classifiers which would allow us to configure the model to predict the valence and arousal values either separately or jointly from a given feature vector. As previously mentioned, predicting both valence and arousal simultaneously proved problematic because there were relatively few training examples in which there was high agreement for both valence and arousal in the same sentence. For the initial phase of the investigation reported in this paper, we chose to predict valence only as the state where there was high agreement among the human raters and had many more training examples.

Even though we collected real-number values for valence and arousal from the AMT workers, we used the MLP as a binary classifier rather than as a regressor. As a result, we took the mean of the raters' valence (x-values) and converted all positive real-number values to '1' and negative values to '0'; these were used as training targets for the classifier. If the binary value of the prediction is 0, it was interpreted as negative valence and if was 1, it was interpreted as positive valence.

4 ARCHITECTURE

Our supplemental emotion pipeline consists of two components, the *Predict* component and the *Dynamical System* component, along

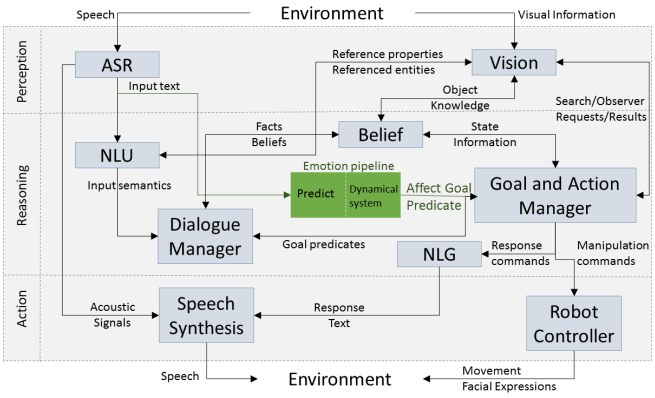

Figure 3: The Emotion Pipeline consists of the prediction and dynamical systems components. It receives utterances from the Automated Speech Recognizer (ASR) and sends its predicted affect to the Goal and Action Manager.

with their incoming and outgoing connections (refer to Figure 3). The Predict component receives a text utterance, extracts its topic proportions and gives these features to the classifier which makes a prediction as to the current emotional state (this corresponds to items 3 and 4 in Figure 1). The prediction is then passed to the Dynamical System component which, at present, simply maps the prediction, $\{0, 1\}$ onto a goal predicate string as $\{0 : frown, 1 : smile\}$. The Dynamical system component sends the goal predicates representing desired affective states to the Goal Manager.

This binary prediction model will eventually be developed to generate finer emotional states. Once that is completed, the Dynamical system component will be developed to represent emotions as 'particles' which have a mass (i.e., the personality of an individual) to which the the predictions apply a 'force' to move the particle using the basic laws of motion. A restoring force, the 'spring', serves to bring the particle to a neutral state. The hypothesis is that modeling emotion as a mass-spring dynamical system will more closely approximate how observed human emotions rise and fall, returning to a neutral state in the absence of input.

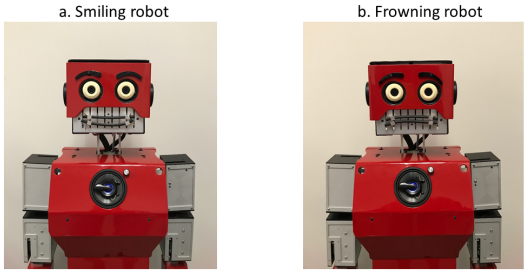

Figure 4: Robo Motio's Reddy robot. a) Smiling b) Frowning

The Goal Manager maintains the systems' beliefs about goals within the world and uses those beliefs to decide upon the proper series of actions to take to reach those goals. Actions the Goal Manger can execute are represented in a script based format. When

an affective goal predicate is submitted to the Goal Manger, it uses an affective control script to select the appropriate *primitive action* which sends messages to the robot controller component to move the robot's actuators to perform pre-specified behaviors.

Here, we are using Robo Motio's Reddy robot (see Figure 4). We defined action primitives for 'smile' and 'frown', which correspond to the robot's facial motors moving to produce a smile or frown, respectively. The action script checks to see if the goal predicate created a 'smile' goal or a 'frown' goal, and performs the proper action primitive. Through this pipeline, therefore, our robot can change its facial expression to match its belief about the world, as predicted by our model.

Using the cognitive robotic architecture to implement the emotion pipeline gives us flexibility in how it can be used and insulates the pipeline from the implementation details of, for example, the robot's affector motors or the type of speech recognizer used. Furthermore, sending the prediction to the Goal Manager gives the Reasoning system an opportunity to coordinate the robot's facial expression with other activities and to include emotion as additional context in the human-robot interaction.

5 DEMONSTRATION

We selected 16 sentences that had not been used to train the model and presented each as textual input to the pipeline using SimSpeech, a speech simulator component, in place of the ASR. We found that the accuracy of the technology currently used in the ASR component was insufficient to adequately transcribe the sentences used in testing. Swapping ASR components did not require any changes to the pipeline implementation, which is an advantage of using the robotic architecture. The test sentences, their ground truth valence and the predicted robot actions, smile or frown, are shown in Table 1. For this test sample, the model correctly predicted the emotional valence in all but two sentences (i.e., numbers 4 and 14), equating to 88% accuracy. Our analysis of why the model makes incorrect predictions is evolving as we gather more experience with different characteristics of our training set. We discuss our thoughts about model performance in the following section.

6 DISCUSSION

The results of the demonstration suggest that the emotion pipeline can be used to appropriately generate a smile or frown expression for the robot, but future work is needed to be able to automatically detect emotional content with more resolution, moving beyond these two coarse levels of valence, negative and positive. Ideally the model could be improved to be able to detect various degrees of positivity or negativity and calm or arousal. That way, the pipeline could be used to inform the emotion dynamics of a conversation as it unfolds between any two participants, such as in therapy. While the relatively high accuracy of 88% seems quite good, further analysis is needed to determine under which circumstances the model makes an incorrect prediction. Preliminary analysis seems to indicate that accuracy is not impacted by sentence word count, but is greatly affected by the number of training examples. The model is also somewhat sensitive the agreement among the human evaluators. We have found that a training set which consisted of sentences in which there was high agreement among the raters was likely to generate a model which was more accurate when given unseen, test sentences. Furthermore, we intend to conduct an HRI (Human-robot Interaction) study to determine to what extent incorrect predictions affect how humans view the validity and usefulness of the robot in this condition.

The method we explored to estimate arousal and valence based on semantics only can, of course, be combined with other components of the robotic architecture, e.g., visual and auditory cues, to get even better estimates, potentially improving the detection of arousal, which we will leave as future work. At present, there seems to be a paucity of experimental results on the accuracy and other statistical measures of validity when inferring human emotion from text transcriptions in the machine learning literature. We shall conduct a review of prior work in area of detecting emotions from text in both the machine learning and sentiment analysis literature to see whether our work can be used to expand the state-of-the-art.

Additionally, our training and testing dataset was comparatively small. We used 448 transcripts to train the LDA model and 953 sentences to train the classifier, a limited amount of data compared to typical machine learning endeavors which may have thousands of training examples available. We are in the process of collecting additional labeled training examples using AMT for the purpose of developing the model's ability to predict additional emotional states beyond positive and negative. Furthermore, in order to generalize the test results to a domain beyond that described in [10], additional training documents from more general domains will have to be used.

7 CONCLUSION

We utilized an automated method we previously created to infer the emotional state of a person with PD and incorporated it into the DIARC robotic cognitive architecture. We view this as a first stage in building an assistive tool which a caregiver could use to infer the emotional state of a person with PD, who lacks an effective facial expressiveness channel. In order for the model to be useful in a clinical setting as a communication assistive tool, it will have to be developed further to generate the more incremental and subtle gradations of human emotional states. In addition, clinical trials would need to be conducted to assess its usefulness.

In our earlier research, we hypothesized that such a tool equipped with the ability to accurately and immediately provide feedback on the emotion content of a conversation is not only beneficial for improving the social interaction with the PD patient, it can improve the quality of life in the home. Since human emotion is communicated via multiple modalities, and through different channels, e.g., voice, facial expressions, gestures, situating such a tool in a robot that appropriately controls the robot's facial motors could compensate for the problematic visual modality of communicating emotion. This will allow us to investigate the extent to which situating such a tool in a robot will enhance its effectiveness and how well the human interacts. The device's technology is not restricted to the domain of PD patients; it should be able to generalize and serve as an intelligent agent useful for monitoring the emotional content of the interaction between any two parties, providing real-time feedback on the emotion valence and arousal as the interaction unfolds.

Table 1: Pipeline demonstration to predict valence using 16 sentences, unseen during model training. Ground truth obtained from humans who rated the content as 1 = postive or 0 = negative emotional valence.

No.	Test Sentence	Ground Truth	Predicted Affect
1	I've fallen and I cant get up.	Neg	frown
2	I am sure that he does not like to shop much	Neg	frown
3	Well, not right now because I cannot do a lot.	Neg	frown
4	Ahh well, we went off to do this and it turn out to be a big ripoff	*Neg*	*smile*
5	Yes, its, uh gotten more, it was, I didn't tell the doctor about it because it came and went.	Neg	frown
6	My frustration is watching my husbands frustration.	Neg	frown
7	Well, I guess a complete lack of mobility	Neg	frown
8	And there isn't much we can do because perception, a persons perception becomes truth to them	Neg	frown
9	Um, I love to read.	Pos	smile
10	The more things that I do on my own instead of having people assist me, I find satisfying.	Pos	smile
11	Well, actually I loved having many children	Pos	smile
12	Im retired for a year now.	Pos	smile
13	I went to a concert yesterday. We took the boat out on the lake out this morning.	Pos	smile
14	But just so when you go in there and you ask people for help some of them are helpful.	*Pos*	*frown*
15	Well, I enjoy driving.	Pos	smile
16	I, what I like doing is physical things like working on my car.	Pos	smile

REFERENCES

[1] Fernando Alonso-Martín, María Malfaz, João Sequeira, Javier F. Gorostiza, and Miguel A. Salichs. 2013. A Multimodal Emotion Detection System during Human–Robot Interaction. *Sensors* 13, 11 (2013), 15549–15581. https://doi.org/10.3390/s131115549

[2] R. A. Calvo and S. D'Mello. 2010. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Transactions on Affective Computing* 1, 1 (Jan 2010), 18–37. https://doi.org/10.1109/T-AFFC.2010.1

[3] F. M. Citron, M. A. Gray, H. D. Critchley, B. S. Weekes, and E. C. Ferstl. 2014. Emotional valence and arousal affect reading in an interactive way: Neuroimaging evidence for an approach-withdrawal framework. *Neuropsychologia* 56, 10 (2014), 79–89.

[4] Scott Deerwester, Susan Dumais, Thomas Landauer, George Furnass, and Laura Beck. 1988. Improving information retrieval with latest semantic indexing. *In: ASIS '88. Information Technology: Planning for the next fifty years. Proceedings of the First Annual Meeting of the American Society for Information Science, Volume 25, Atlanta, Georgia, 23-27 October 1988 Edited by Christine L. Borgman and Edward Y. H. Pai* (10 1988).

[5] E. DeGroat, K. D. Lyons, and L. Tickle-Degnen. 2006. Verbal content during favorite activity interview as a window into the identity of people with Parkinson's disease. *Occupational Therapy Journal of Research: Occupation, Participation, and Health* 26, 2 (2006).

[6] C. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109. In *International AAAI Conference on Web and Social Media*.

[7] Jonathan Posner, J.A. Russell, and B. S. Peterson. 2005. "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology". *Development and Psychopathology* 17 (2005), 715–734. https://doi.org/10.1017/s0954579405050340

[8] Matthias Scheutz, Thomas Williams, Evan Krause, Bradley Oosterveld, Vasanth Sarathy, and Tyler Frasca. 2019. An overview of the distributed integrated cognition affect and reflection DIARC architecture. In *Cognitive Architectures*. Springer, 165–193.

[9] H. Tao and T. Tan. 2005. Affective computing: A review. *International Conference of Affective Computing and Intelligent Interaction* (2005), 981–995.

[10] L. Tickle-Degnen, T.D. Ellis, M. Saint-Hilaire, C. Thomas, and R. C. Wagenaar. 2010. Self-management rehabilitation and health-related quality of life in Parkinson's disease: A randomized controlled trial. *Movement Disorders* 25 (2010), 194–204.

[11] A. Valenti, M. Chita-Tegmark, A. Bock, and M. Scheutz. 2019b. In their own words: Using topic modeling to describe the emotional state of persons with Parkinson's disease. (2019b). submitted.

[12] A. Valenti, M. Chita-Tegmark, L. Tickle-Degnen, A. Bock, and M. Scheutz. 2019a. Using topic modeling to infer the emotional state of people living with Parkinson's disease. (2019a). under review.

[13] Li Zhang, Ming Jiang, Dewan Farid, and M.A. Hossain. 2013. Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot. *Expert Systems with Applications* 40, 13 (2013), 5160 – 5168. https://doi.org/10.1016/j.eswa.2013.03.016