Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5739–5748
Marseille, 20-25 June 2022

HADREB: Human Appraisals and (English) Descriptions of Robot Emotional Behaviors

Josue Torres-Fonseca, Casey Kennington
Boise State University
1910 W University DR, Boise, ID 83725
josuetorresfonse@u.boisestate.edu
caseykennington@boisestate.edu

Abstract

Humans sometimes anthropomorphize everyday objects, but especially robots that have human-like qualities and that are often able to interact with and respond to humans in ways that other objects cannot. Humans especially attribute emotion to robot behaviors, partly because humans often use and interpret emotions when interacting with other humans, and they apply that capability when interacting with robots. Moreover, emotions are a fundamental part of the human language system and emotions are used as scaffolding for language learning, making them an integral part of language learning and meaning. However, there are very few datasets that explore how humans perceive the emotional states of robots and how emotional behaviors relate to human language. To address this gap we have collected HADREB, a dataset of human appraisals and English descriptions of robot emotional behaviors collected from over 30 participants. These descriptions and human emotion appraisals are collected using the Mistyrobotics Misty II and the Digital Dream Labs Cozmo (formerly Anki) robots. The dataset contains English descriptions and emotion appraisals of more than 500 descriptions and graded valence labels of 8 emotion pairs for each behavior and each robot. In this paper we describe the process of collecting and cleaning the data, give a general analysis of the data, and evaluate the usefulness of the dataset in two experiments, one using a language model to map descriptions to emotions, the other maps robot behaviors to emotions. **Keywords:** emotion, human-robot interaction, language learning

1. Introduction

We present *Human Appraisals and Descriptions of Robot Emotional Behaviors* (HADREB), a dataset of human appraisals and descriptions of over 1,000 robot emotional behaviors. HADREB represents a rich resource for investigating how humans perceive robotic behaviors relating to their emotional states.[1] HADREB will be useful for incorporating emotion into language models and generating novel behaviors that allow robots to express specific emotions. This dataset is especially useful in solving the cold-start problem of spoken dialogue systems that have no prior knowledge of language (McNeill and Kennington, 2020) because emotion exists in humans before they learn language. Emotion is also tied to the meanings of many words (Lane and Nadel, 2002); in particular, abstract words (e.g., *democracy* and *utopia*) are grounded in emotion (Vigliocco et al., 2013).

Novikova et al. (2015) showed that humans anthropomorphize robots by attributing emotional content to robot behaviors, no matter how simple those behaviors might be, and no matter how the robot looks. Taking inspiration from Moseley et al. (2012) which posits that emotion is tied to the motor system, we design this dataset to leverage the fact that humans anthropomorphize robot behaviors for emotional content by collecting human appraisals of observed robot behaviors, and we collect the corresponding behaviors—the robotic motor system—and descriptions to bring together the modalities of emotion, behaviors, and language.

This dataset builds on prior work, most directly McNeill and Kennington (2019), that identified how three different modalities (1) robot behaviors represented as internal states, (2) robot faces, and (3) robot sounds contribute to different emotions. However, this dataset differs from that work in several ways. Whereas their work used 16 emotion labels, ours separates emotion labels into 8 positive and negative valence pairs of the 16 emotions, and participants were able to assign a graded Likert-style label for each pair (similar to Experiment 3 of their work). These 8 valence pairs were identified by Robinson (2008) as being a taxonomy of common emotions. Furthermore, our dataset includes data that will enable future modeling of robot behaviors tied to emotion and descriptions, whereas prior work had no way of knowing how the behaviors were invoked. Moreover, and more importantly, our dataset was collected from in-person participants using two robots rather than from online participants and recordings of a single robot. This is crucial because in-person, co-located communication is the setting for language learning in humans (Fillmore, 1981; McCune, 2008). This highlights the fact that this kind of data is very challenging to collect as it requires multiple robot platforms and in-person human participants.

In the following section, we describe related work comparing our dataset with other similar datasets. We then explain how we collected and processed the data, then we offer some analysis of the data. We then perform two experiments, following Moro et al. (2020), to test the mapping between a representation of the robotic be-

[1] We use the term *emotion* rather than *affect* because emotion is more tied to the linguistic system, whereas affect is a more basic and shorter-term sense of feeling (Barrett, 2017).

Figure 1: Cozmo (left) and Misty II (right) Robot Platforms

haviors to the emotion labels, as well as the descriptions of those behaviors to the emotion labels. Our results compare well to prior work, with nuanced differences that open the door for further research in natural language processing and human-robot interaction tasks.

2. Related Work

Compared to other language datasets used as language resources, our dataset aims to provide insight on how people interpret the emotional display of robots as they perform randomly generated behaviors. This expands on the work done in (McNeill and Kennington, 2019), and allows our model to be used in incorporating emotion to language models possibly improving them and/or interactions between robots and humans. Recently comparable data collection efforts are relatively rare. Examples include Fan et al. (2017) and Chita-Tegmark et al. (2019) who explore how humans perceive the emotional intelligence of robots, but offer no dataset tied to descriptions of those robot behaviors. Pena and Tanaka (2020) analyzed human perception of a social robot's emotional states via facial and thermal expression and Spatola and Wudarczyk (2021) attempt to understand how humans anthropomorphize robots, but neither offer a dataset that brings together robot behaviors, descriptions, and emotion labels as we do here. Despite the rarity of similar datasets, other datasets have been gathered relating to emotion in virtual agents. Much of these datasets relate to analyzing how persons perceive generated facial expressions (Beer et al., 2015; Randhavane et al., 2019; Beer et al., 2009), whereas here we focus on physical robots and capture not only facial expressions of the robots, but also their movements and sounds which have implications for how humans judge emotional displays (McNeill and Kennington, 2019).

Relating to the importance of analyzing how people perceive the emotions of robots/virtual agents, researchers have also expressed the importance of understanding human emotions. Alternatively, there are other datasets about human emotions such as in Kosti et al. (2017), which presents a dataset about human emotions in real environments. Others are more closely related to human robot interaction such as in Jam et al. (2021), which proposes a data-driven method for increasing the number of emotion classes present in human-robot interactions.

Related to language models is the task of analyzing and attempting to classify emotion in text, which is especially prevalent in the field of natural language processing, often referred to as sentiment analysis (Poria et al., 2018; Demszky et al., 2020; Liu et al., 2019). Our work differs in that we are not attempting to recover perceived emotional states of text writers, rather we are attempting to tie emotion to language via robot behaviors. As robots do not actually feel emotions, their behaviors are a useful approximation for emotional states, as behaviors are in humans when people interpret the emotional state of others; i.e., motor system plays a role in abstract emotional meaning (Moseley et al., 2012).

Finally, datasets have also been gathered concerning emotion in speech. Due to difficulties of collecting natural scenarios consisting of spontaneous emotions much of the dialogue/speech is often collected from actors given emotions to convey (Rambabu et al., 2020), (Asai et al., 2020; Cao et al., 2014; Livingstone and Russo, 2018). Others attempt to make the dialogue more natural by giving actors scenarios meant to elicit particular emotions without giving specific emotion labels (Busso et al., 2016). Overall, the goal of these datasets are to produce dialogue which express certain emotions. In contrast, we are not gathering emotional speech dialogue from human participants, instead robot behaviors "express emotion" as perceived by the participants through a speech-synthesizer. Furthermore, the robots do not participate in dialogue; rather they are limited to short utterances stated as either a question or exclamation.

3. Robots: Cozmo and Misty

To collect data about human appraisals and descriptions of robot emotional behaviors, we used the Digital Dream Labs Cozmo and Mistyrobotics Misty II robots (see Figure 1). This section describes both robot platforms.

Misty Misty has a height, depth, width and weight of 35.56 cm, 25.4 cm, 20.32 cm and 2.7 kgs respectively. With respect to hardware used for this dataset it has:

- two arms with 119 degrees of freedom
- 2 high-fidelity speakers
- moveable neck (pitch, yaw, and roll)
- 4" LCD image display/screen

Cozmo Cozmo has base dimension of 10.795 cm (length) by 6.35 cm (width) by 10.16 cm (height with lift at maximum height). Hardware-wise Cozmo is equipped with the following:

- a lift for small objects
- track driving tread system
- a small OLED display screen for the face
- speaker for speech synthesis

Figure 2: Sample of three Cozmo faces.

4. Data Collection

Our goal was to collect data that would allow researchers to map from descriptions or representations of short, robot behaviors to emotional appraisals of those behaviors. In this section, we describe how we generated behaviors on both robotic platforms, the procedure for human participants labeling the behaviors physically in-person with the robots, and analyses of the data we collected.

4.1. Generating Random Behaviors

To begin work on data collection we first generated our own custom behaviors that participants could describe and assign a graded valence (i.e., positive or negative emotion) label. We explain below how we generated novel behaviors and used a Likert-style labeling process for emotion appraisals and descriptions of robot behaviors.

Generating Actions For generating actions, we made use of various movements, facial configurations, and sounds for both Cozmo and Misty. For Cozmo we generated actions by making use of head and lift position and left and right wheel speed/direction. In Misty's case, we generated actions by making use of arm and head position (we did not allow Misty to move its wheels as it is much larger than Cozmo and would easily fall from the table). For both robots, any action using any of these movements requires a parameter to invoke that movement. However, Cozmo required a duration parameter of how long the robot should take to change its state to match that parameter, while Misty required a velocity parameter of how quickly Misty should move its arm and head. We constrained the possible parameter values for each of these movements to fall within a range of values that we randomly sampled, as shown in the Appendix.

Generating faces To generate faces, we gave Cozmo a set of 13 possible face images, recreated from existing examples. Misty was programmed with a set of 11 possible face images, which we used. For each behavior, the program randomly chooses one of these images to display on Cozmo or Misty's OLED display for the duration of the behavior. Examples of three Cozmo and Misty faces can be seen respectively in Figures 2 and 3.

Generating sounds For sounds, we constructed a list of vowel sound approximations stated as either a question or exclamation for a total of 14 utterance options for Cozmo:

Figure 3: Sample of three Misty faces. These faces are provided with the Misty platform.

- 'ehhhh?' and 'ehhhh!'
- 'aa?' and 'aa!'
- 'uu?' and 'uu!'
- ''rue?', and 'rue!'
- 'eyy?' and 'eyy!'
- 'oh!'
- 'hm'
- 'oi'
- 'umm'

While Misty had 6 utterances options:

- 'oh!'
- 'hmm'
- 'oi'
- 'umm'
- 'aa?' and 'aa!'

For each behavior, we randomly selected one of these sounds and passed it through the robot's built in speech synthesizer (trained for English) to play for the duration of the animation. Cozmo's speech synthesizer has a young-sounding voice, whereas Misty's voice is closer to an adult voice.

Final Behaviors Taking the action, face, and utterance generation procedures together, we generated novel random behaviors. For all behaviors there was a 50% chance of an action, face or utterance being generated and we randomized behavior length by randomly repeating the process above 1-5 times, forming a random sequence of short actions into a 4-10 second behavior.

5. Participant Procedure

For purposes of data collection, we recruited participants from Boise State University to observe as many novel behaviors as possible in an hour.[2] Participants were paid a nominal payment for their participation. Sessions were split so the participant spent 30 minutes with Misty and 30 minutes with Cozmo. We invited the participants into our lab where the Cozmo robot was on one table near a laptop, and the Misty robot was on another table near a different laptop. After reading and agreeing to an informed consent, the experimenter explained that the participant was to use the website displayed on the laptop. After clicking *Start* the participant was taken to a page that explained that they needed to watch the robot so they could describe

[2]Data was also gathered by placing robots in public areas. Only a handful of data entries were gathered this way and no information about these participants is included in this paper.

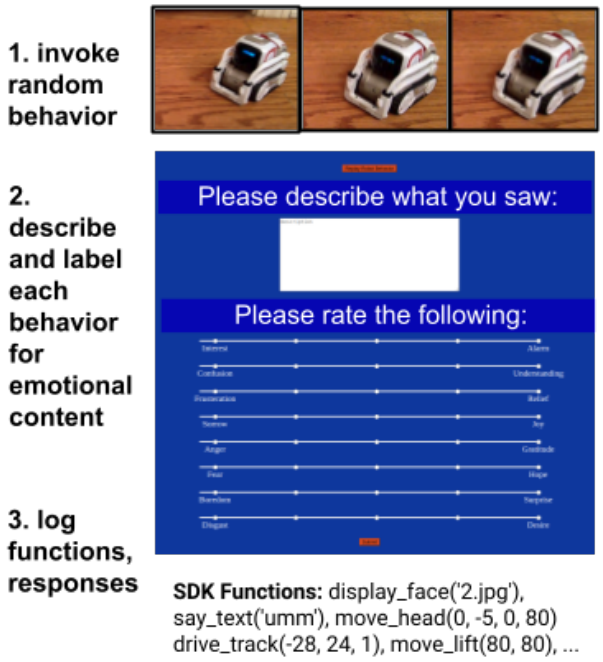

Figure 4: Process of collecting data from the perspective of a participant with example of logged functions, descriptions, and emotion ratings.

positive valence	negative valence
interest	alarm
understanding	confusion
relief	frustration
joy	sorrow
gratitude	anger
hope	fear
surprise	boredom
desire	disgust

Table 1: Valence of 16 specific emotions.

its behavior and its emotions. When ready, the participant clicked on a button that invoked a behavior, then they were automatically taken to the annotation page where they typed a description in a text input box and used radio buttons for the Likert ratings (1-5) of each of the 8 valence pairs shown in Table 2. The participants could optionally click on a button on the top of the page to re-invoke the exact behavior that they saw as many times as needed. When they were done, they clicked on a *Finished* button, and they were taken back to the start page where they repeated the process. The software logged the responses and set of robot functions/parameters used to generate the behavior. This process is depicted in Figure-4.

Following the format of the Godspeed Questionnaire (Bartneck et al., 2009), we displayed the negative valence on the left and the positive valence on the right with a 5-point scale between them. The participants were required to score all of the valence pairs for all observed behaviors. The collection of the data was facilitated through the use of a web survey application (see Figure 4), which the participants used to invoke behaviors and give their analysis of the behaviors per given instructions. The web application stored the descriptions and emotion labels given in a CSV file, while the functions, which were used to produce a behavior, were stored in a text file. In both the function and questionnaire results files we paired the data with timestamps to act as an ID number allowing identification of what behavior (represented as functions) were being described and analyzed by a participant. In total, we recruited 32 participants (23 male, 9 female).[3]

Representing Robot behaviors Following the data collection phase, we then gathered the internal state data, facial images, and sounds of the behaviors for both Misty and Cozmo. Defining internal states and the process for collecting them is robot platform-dependent. For Cozmo, the states are represented as a vector of state properties, and for Misty they are represented as robot state variables that one can subscribe to. For both, a change in any state property or variable results in a state change update which is logged, with timestamps. We identified internal state properties and variables that we thought would be useful to the research community; the logged functions (and parameters) are available for researchers to invoke if the need arises for deriving other information from the exact behaviors that the participants observed.

The variables and properties for Misty and (a sample) for Cozmo are listed below:

Misty:

- head pitch, roll and yaw angles and velocity
- left arm and right arm angles
- x, y and z acceleration values

Cozmo:

- lift angle and height
- head angle
- left and right wheel speed
- if picked up
- cliff detected
- has in progress actions
- ...

In total 11 and 43 variables and properties were collected for Misty and Cozmo Respectively.

5.1. Data Processing

Following data collection, we cleaned the data using 4 processing techniques to allow for easier use and ac-

[3]We inferred the number of males and females from participant names.

cess. We describe the data processing below.

a) Duplicate IDs: Due to a bug in the web application used to collect the data, duplicate data entries were logged when participants clicked the submit button more than once. To alleviate any data issues, we removed duplicate entries through identification of duplicate IDs. This resulted in the removal of 10 data entries for the data collected for Cozmo and 27 data entries for that of Misty.

b) No behavior: When reviewing the data, we discovered that there were multiple entries where the participant stated that the robot "did nothing" or that "it glitched out." This was likely caused due to how Cozmo runs programs through communication with a phone (connected to a WiFi signal on Cozmo) where the phone collects instructions from a program on a computer (the phone is connected to) and then those instructions are sent to Cozmo from the phone thereby allowing for communication with the program to fail in multiple places. To identify descriptions which indicated a failure in performing a behavior, we identified data entries where all labels were the same, as participants were instructed that they may label all emotions with a 3 if they were confused. Then we manually scrolled through the descriptions to find entries where failure in performing a behavior was described. Due to the nature of Cozmo's communication, this issue more commonly occurred when participants used Cozmo resulting in 12 entries being removed this way for Cozmo. For Misty one entry was removed this way indicating that this may have been a one-time error.

c) Empty entries: Despite existing checks for empty descriptions or emotion values in the web application we found that there was one entry in the Cozmo data with an empty description. This entry was unusable and was removed.

d) Lack of Internal State Data: As described in the data collection section, using timestamps as IDs, we gathered the internal state data of each behavior and attached them to the appropriate description and emotion labels. However, as we were organizing the data we discovered that there were many entries where either the description and emotion label did not have corresponding internal state data or vice versa. Since all of this data is needed together in order for the entries to be useful these were removed. This led to the removal of 82 and 99 entries for Cozmo and Misty respectively.

6. Data Analysis

Once the data cleansing phase was completed, a total of 547 and 545 complete entries remained for Cozmo and Misty respectively. In addition to cleaning the data we analyzed the data set to find patterns and interesting new perspectives on the raw data.

6.1. Emotion Labels

Part of the analysis includes a mean and standard deviation of all the emotion labels for both Misty and Cozmo which can be seen in Table 2. Looking at the data it is clear that most emotion labels for both Misty and Cozmo stay between a 2 and 4 centering around a label of 3, which can be interpreted as neutral or no emotion. This indicates that most participants observing Cozmo and Misty's emotional state observed the robots as being, overall, neutral in most emotions though individual behaviors have some emotions as deviating from neutral. This indicates to us that there is no particular bias in the data towards any individual emotion.

Emotion Pair	Cozmo Mean (STD)	Misty Mean (STD)
Interest/Alarm	2.39 (1.04)	2.7 (1.24)
Confusion/Underst.	2.78 (0.99)	2.86 (1.74)
Frustration/Relief	2.92 (0.77)	2.92 (0.82)
Sorrow/Joy	2.9 (0.86)	2.96 (0.91)
Anger/Gratitude	2.96 (0.7)	2.96 (0.83)
Fear/Hope	3.07 (0.75)	2.99 (0.81)
Boredom/Surprise	3.08 (0.82)	3.18 (1.09)
Disgust/Desire	3.13 (0.84)	3.05 (0.86)

Table 2: Means and Standard deviations (STD) for participant responses for the 8 valence pairs.

Another informative observation from the analysis shows that for both Cozmo and Misty *Boredom/Surprise*, *Disgust/Desire*, and *Interest/Alarm* tended to lean towards the positive emotion. This indicates that both machines are capable of showing these emotions with the actions they are able to perform. Since both are able to show these emotions, the analyses indicate that they are most likely representative of actions shared by both Cozmo and Misty. In the case of Cozmo and Misty, the only actions they both share are the ability to change their facial display and the ability to speak using their built in speech-synthesizer. Therefore, indicating that the emotions of *Surprise*, *Desire* and *Alarm* likely involve the use of face and audio actions (e.g. widening of the eyes and exclaiming to express surprise), as reported in McNeill and Kennington (2019).

6.2. Descriptions

In addition to analyzing the emotion labels, we also analyzed the descriptions for Cozmo and Misty's behaviors. We calculated the number of total tokens, average description length, and size of vocabulary. For Cozmo, the total number of tokens were 8,349 tokens, average description length was 15.26 words, and vocab size was 622 unique tokens. For Misty, the total number of tokens were 9,953 tokens, average description length was 18.26 words and vocab size was 866 unique tokens. It is important to note that vocab size was gathered after preprocessing of the descriptions involving lemmatization and lowercasing of the tokens as well as filtering out non alphabetic tokens. Therefore, from this data it is clear that participants tended to have more observations to give about Misty than they

did for Cozmo. This could be due to the fact that Misty has a higher variability of actions available to perform compared to Cozmo, as Misty can move its head in accordance with pitch, yaw, and roll and move it's left and right arms while Cozmo can only move it's head in accordance with pitch and move one lift (though Cozmo could move using its wheels). This argument makes sense as descriptions did involve describing what the robot did, not their perceived intent of the robot. However, in many of the descriptions, in addition to describing what the robot did, participants also tended to describe what emotions these behaviors were displaying. Therefore, this seems to indicate that robots with more variability in actions and movement abilities tend to be able to express more emotions as well as better express those emotions so that an observer is more confident in what emotion a robot is showing. This is supported by Misty's means and higher standard deviation values for emotion labels as compared to Cozmo seen in Table 2.

Other than looking at the length of descriptions, we analyzed word frequency of words in the descriptions, ignoring stop words, for both Cozmo and Misty seen in Figure 5.

The data offers interesting insights on what participants thought was most important to describe in the robots behavior or what stood out to the participants. For Misty, the top 5 words were *eye*, *head*, *arm*, *moved* and *said*, words that denote salient physical aspects of the robot and common verbs associated with robot movements. On the other hand, for Cozmo, the top 5 descriptive words were *moved*, *arm*, *head*, *forward*, and *raised*. In both, top 5 word lists *moved*, a generic verb, is part of the list and in Cozmo *forward* and *raised* were also included. This indicates that for most participants movement was a crucial part of what was observed to interpret a robot's emotion. This follows with how humans tend to use gestures or facial movements to express their emotional state. This is further supported by how in both word frequency lists *arm* and *head* are present, which are parts of the robot for both Cozmo and Misty that are able to move. Interesting to note is that in Misty's word frequency list *head* is more frequent than *arm* and for Cozmo *arm* is more frequent than *head*. This is likely due to how for Misty the head has more degrees of freedom than the arm, though for Cozmo the arm has more range of motion than the head. This further shows that motion is an important part of expressing emotion understandable by humans. For Misty *eye* and *said* were also high in the list. However, for Cozmo anything related to the face display or speaking was farther down in the list. Specifically, *said* and *looked*, which were 9 and 10 on the list respectively. We conjecture that this follows from prior research that Cozmo is viewed as a young child, therefore sounds are not interpreted as uttered words (Plane et al., 2018). Moreover, the more frequent use of *eye* for Misty can be explained by how Misty has a higher variation of facial configurations than Cozmo. The fact that *eye* was the most frequent word in Misty's description also shows how important using eyes are to expressing emotion.

6.3. Internal State Data

Analysis of the internal state data included a visualization of the distribution of possible value ranges for each variable (see the Appendix). For most internal state values, the values follow a normal distribution curve or the values are very one sided (e.g. Actuator Left Arm). This is expected given how the behaviors were generated. Even though the behaviors were randomized in what actions were included and for some how the actions were performed (e.g. head pitch), some values were set to a hard number as too much randomization would be excessive. This limitation is made very clear in Cozmo's visualization of the distribution of possible ranges of its state variables. Furthermore, many of these variables were set so they wouldn't be too subtle otherwise the participant would not have much of an action to observe. Therefore, even though they were randomized we limited the range of these variables so the actions were clear enough for the observer. Importantly, actions were often a small part of a larger sequence of actions that composed the entire observed behavior; even similar small actions in a different sequence results in a very different behavior.

7. Experiments

The goal of our experiments is to showcase the utility of the dataset—we are not proposing any new models or methods. For both of our experiments, we follow Moro et al. (2020) (which improved upon the model proposed in McNeill and Kennington (2019)) to first show that a transformer language model can reliably map from descriptions to emotion labels in a zero-shot learning task, and second to show that a simple model using specific features can map from a representation of the robot internal states to the emotion labels.

7.1. Experiment 1: Descriptions to Emotions

In this experiment, we use the BART model (Lewis et al., 2019) for a zero-shot learning task.

Task, Procedure, Metrics As we are using zero-shot learning, we are not training or fine-tuning. The BART model is pre-trained on English data. The architecture is a model with a bidirectional encoder and auto-aggressive decoder, effectively benefiting from BERT and GPT-like models. BART uses an arbitrary noising function to corrupt text, then the decoder reconstructs the original. We use the huggingface (Wolf et al., 2020) *bart-large-mnli* model in a zero-shot classification pipeline. Such a pipeline takes text as input and produces a probability distribution over specified labels. The labels we specify are the names of the 16 emotions listed in Table 2.

We use all of the data collected for Cozmo and Misty to evaluate. For each description, we used the BART

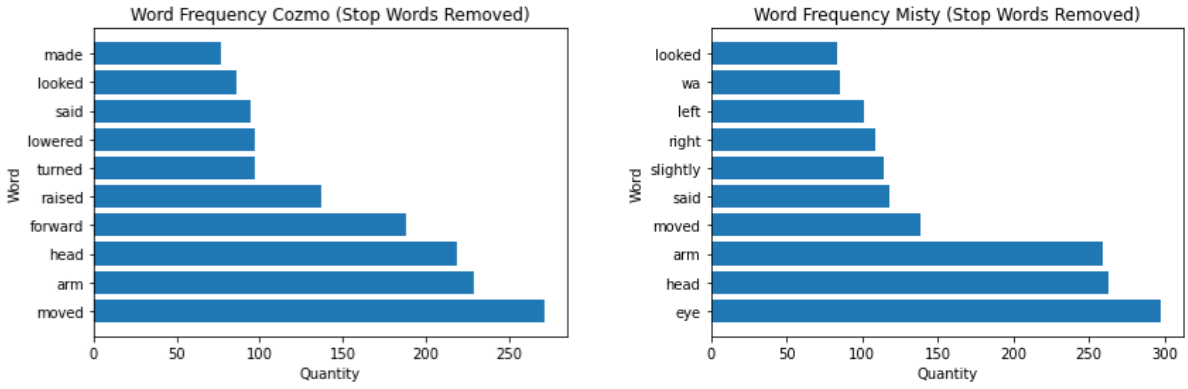

Figure 5: Left: Cozmo Word Frequency Right: Misty Word Frequency

pipeline to produce a distribution over the 16 emotion labels. We then took the label with the highest probability and compared it to the emotion labels. As the emotion labels were graded Likert values (1-5, 1 being the negative emotion and 5 being the positive emotion) between 8 valence pairs, we assigned a full accuracy point if the label was 1 (for the negative emotion) or 5 (for the positive emotion), and a half a point of the label was 2 (for the negative emotion) or 4 (for the positive emotion). Cases where all of the emotions were labeled with 3 (i.e., the participant did not think the robot was showing any particular emotion) were ignored. This resulted in 333 (out of 545) instances for Misty and 238 (out of 547) for Cozmo. The accuracy is therefore somewhat inflated, but the metric is still useful to give a starting point for how this data could be used. We leave more principled experiments and evaluations using the descriptions for future work.

Results The Misty data yielded a **79.3%** accuracy, and the Cozmo data yielded a **69.5%** accuracy. These are lower than the metrics reported in Moro et al. (2020), but that work only used Cozmo. It is unsurprising that the Misty data has higher accuracy as Misty did not move its wheels and was easier to describe. Moreover, Misty's face showed obvious emotion and emotion words were often used in the descriptions making the zero-shot learning task easier for the model. These results do showcase the usefulness of the descriptions and emotion labels.

7.2. Experiment 2: Robotic States to Emotions

The model in this experiment uses internal robot states only, and maps them to a set of features derived from Novikova et al. (2015) termed *Novikova* features. Following the descriptions in Moro et al. (2020), we use 9 of the proposed features described below:

- Approach 1 - Transfer weight forward (head bent or movement forward)
- Approach 3 - Move its body forward (track wheel movement forward)
- Approach 5 - Extend or expand its body (lift movement up)
- Avoidance 6 - Transfer weight backward (head bent or movement backward)
- Avoidance 8 - Move its body backward (track wheel movement backward)
- Avoidance 9 - Attract limbs close to body (lift movement down)
- Energy 11 - High strength (high wheel speed)
- Energy 12 - Low strength (low wheel speed)
- Flow 18 - High change in tempo (change in motor speed)

Each feature yields a value that is a percentage of the time that feature is true for the duration of the behavior. For example, Approach 1 is the percentage of robot state changes with forward movement and Avoidance 9 is the percentage of the state updates where the lift was in the lower half. Taken together, these transformations result in a vector of 9 values, each value normalized between 0 and 1.[4]

Moro et al. (2020) mentions that the Novikova features allow for potential generalizability across multiple robot platforms because one only needs to map the internal state representations of a chosen robotic platform to the Novikova features, though they did not actually evaluate on multiple robot platforms. Here, we test this hypothesis of generalizability because we have collected data from two robot platforms.

Metrics, Task & Procedure We first consider the Cozmo and Misty datasets in isolation. For each, we randomly select 100 test samples and train on the remaining (447 for Cozmo and 445 for Misty). We compute the f1 score and accuracy of the model correctly predicting that an emotion was selected above a Likert value of 3 (out of 5, 5 representing the positive side of the valence pair) for each of the 8 valence paris (note that this means each behavior will result in 8 possible values, each contributing to an overall f1 score and accuracy).

[4]The scripts we used to transform the features from Cozmo and Misty into Novikova features are included as part of the dataset.

Setting	f1 score	accuracy
Cozmo train/eval	58.29	79.61
Misty train/eval	53.51	74.86
Cozmo train, Misty eval	54.36	76.35
Misty train, Cozmo eval	51.27	69.12
Moro et al. (2020)	57.0	91.0

Table 3: Experiment 2 Results: Train and evaluations on Comzo, Misty, and across both datasets using the Novikova features.

We then consider both datasets and if the model and Novikova features can generalize by first training on all of the Cozmo data (547 behaviors) and evaluating on the Misty data (545 behaviors), then training on the Misty data and evaluating on the Cozmo data.

Model & Training Using the same model as Moro et al. (2020) (Experiment 2), we use the multinomial K-Nearest Neighbor classifier (Zhang and Zhou, 2007). The only parameter that was needed for the KNN classifier was number of neighbors, which we set to 5 to balance generalizability and performance in our task. Even though our results are not directly comparable to Moro et al. (2020) because they used a different dataset using just the built-in behaviors of Cozmo, we nonetheless make a comparison and discuss the implications of the data and model.

Results Table 3 shows the results for this experiment. We note that the results are substantially lower for the accuracy metric then those reported in Moro et al. (2020) using the same model, though the results for the f1 score metric are comparable. We take this to mean that the model is generalizable when using the Novikova features (though note that the two platforms have many similarities such as track wheels, OLED faces, and simplistic arms/lift making the generalizability claim only partially substantiated). We note that when training on Cozmo data and evaluationg on Misty data, the results are better because Cozmo uses wheel movements which map to some of the Novikova features, whereas Misty was not allowed to move its wheels; therefore a model trained on the Misty data could not learn about movement features. Still, the a model can reliably learn using the data we collected, substantiating the claim that the data can be useful for ongoing research. The simplicity of the model opens the door to more nuanced approaches to this and other tasks using this dataset.

8. Conclusion

We have presented HADREB, a dataset of human appraisals and descriptions of robot emotional behaviors. The dataset contains English descriptions and appraisals with more than 1,000 total descriptions and graded valence labels of eight emotion pairs for each behavior. The analysis of this dataset provides insight on how the abilities of the two robots and their differences affect the perceived 'emotional' states of these robots for human observers. Though given the small size of the dataset, the analysis may be limited due to lack of results that may be present in larger datasets of the same kind. We also evaluated the usefulness of this dataset by a zero-shot learning task that mapped from descriptions to emotion labels fairly reliably, and by using the data in a model that classifies emotion labels from the internal states of robot behaviors. Results from the experiments were respectable, though limited as an important starting point. Overall, the HADREB dataset provides a rich resource for human perception of robots' emotional states in human-robot interaction. This is especially true, given the in-person, co-located environment the data was collected, which attempts to mimic the setting for language learning in humans. We plan on future releases of the dataset with more data. We hope that feedback from the community will inform us as to what we might change about future data collections. HADREB is publicly available to the community.[5]

Acknowledgements We are grateful to the anonymous reviewers for their helpful feedback. We also acknowledge Josh Coward who helped develop the web interface for the data collection tool.

9. References

Asai, S., Yoshino, K., Shinagawa, S., Sakti, S., and Nakamura, S. (2020). Emotional speech corpus for persuasive dialogue system. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 491–497.

Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1):1–23, jan.

Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1(1):71–81.

Beer, J. M., Fisk, A. D., and Rogers, W. A. (2009). Emotion recognition of virtual agents facial expressions: the effects of age and emotion intensity. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 53, pages 131–135. SAGE Publications Sage CA: Los Angeles, CA.

Beer, J. M., Smarr, C.-A., Fisk, A. D., and Rogers, W. A. (2015). Younger and older users [U+05F3] recognition of virtual agent facial expressions. *International journal of human-computer studies*, 75:1–20.

Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi, N., and Provost, E. M. (2016). Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80.

[5] https://github.com/bsu-slim/hadreb

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390.

Chita-Tegmark, M., Lohani, M., and Scheutz, M. (2019). Gender effects in perceptions of robots and humans with varying emotional intelligence. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 230–238. IEEE.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A. S., Nemade, G., and Ravi, S. (2020). Goemotions: A dataset of fine-grained emotions. *CoRR*, abs/2005.00547.

Fan, L., Scheutz, M., Lohani, M., McCoy, M., and Stokes, C. (2017). Do we need emotionally intelligent artificial agents? first results of human perceptions of emotional intelligence in humans compared to robots. In *International Conference on Intelligent Virtual Agents*, pages 129–141. Springer.

Fillmore, C. J. (1981). Pragmatics and the description of discourse. *Radical pragmatics*, pages 143–166.

Jam, G. S., Rhim, J., and Lim, A. (2021). Developing a data-driven categorical taxonomy of emotional expressions in real world human robot interactions. *CoRR*, abs/2103.04262.

Kosti, R., Alvarez, J. M., Recasens, A., and Lapedriza, A. (2017). Emotic: Emotions in context dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2309–2317.

Lane, R. D. and Nadel, L. (2002). *Cognitive Neuroscience of Emotion*. Oxford University Press.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Liu, C., Osama, M., and de Andrade, A. (2019). DENS: A dataset for multi-class emotion analysis. *CoRR*, abs/1910.11769.

Livingstone, S. R. and Russo, F. A. (2018). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

McCune, L. (2008). *How Children Learn to Learn Language*. Oxford University Press.

McNeill, D. and Kennington, C. (2019). Predicting human interpretations of affect and valence in a social robot. In *Robotics: Science and Systems*.

McNeill, D. and Kennington, C. (2020). Learning word groundings from humans facilitated by robot emotional displays. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 97–106, 1st virtual meeting, July. Association for Computational Linguistics.

Moro, D., Caracas, G., McNeill, D., and Kennington, C. (2020). Semantics with feeling: Emotions for abstract embedding, affect for concrete grounding.

Moseley, R., Carota, F., Hauk, O., Mohr, B., and Pulvermuïler, F. (2012). A role for the motor system in binding abstract emotional meaning. *Cereb. Cortex*, pages 1634–1647, July.

Novikova, J., Ren, G., and Watts, L. (2015). It's Not the Way You Look, It's How You Move: Validating a General Scheme for Robot Affective Behaviour. In Julio Abascal, et al., editors, *Human-Computer Interaction – INTERACT 2015*, pages 239–258, Cham. Springer International Publishing.

Pena, D. and Tanaka, F. (2020). Human perception of social robot's emotional states via facial and thermal expressions. *ACM Transactions on Human-Robot Interaction (THRI)*, 9(4):1–19.

Plane, S., Marvasti, A., Egan, T., and Kennington, C. (2018). Predicting perceived age: Both language ability and appearance are important. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 130–139, Melbourne, Australia, July. Association for Computational Linguistics.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2018). MELD: A multimodal multi-party dataset for emotion recognition in conversations. *CoRR*, abs/1810.02508.

Rambabu, B., Botsa, K. K., Paidi, G., and Gangashetty, S. V. (2020). Iiit-h temd semi-natural emotional speech database from professional actors and non-actors. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1538–1545.

Randhavane, T., Bera, A., Kapsaskis, K., Sheth, R., Gray, K., and Manocha, D. (2019). Eva: Generating emotional behavior of virtual agents using expressive features of gait and gaze. In *ACM symposium on applied perception 2019*, pages 1–10.

Robinson, D. L. (2008). Brain function, emotional experience and personality. *Netherlands Journal of Psychology*.

Spatola, N. and Wudarczyk, O. A. (2021). Ascribing emotions to robots: Explicit and implicit attribution of emotions and perceived robot anthropomorphism. *Computers in Human Behavior*, 124:106934.

Vigliocco, G., Kousta, S.-T., Della Rosa, P. A., Vinson, D. P., Tettamanti, M., Devlin, J. T., and Cappa, S. F. (2013). The Neural Representation of Abstract Words: The Role of Emotion. *Cerebral Cortex*, 24(7):1767–1777, 02.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020).

Huggingface's transformers: State-of-the-art natural language processing.

Zhang, M.-L. and Zhou, Z.-H. (2007). Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048.

A. Sample of Robot Internal State Ranges

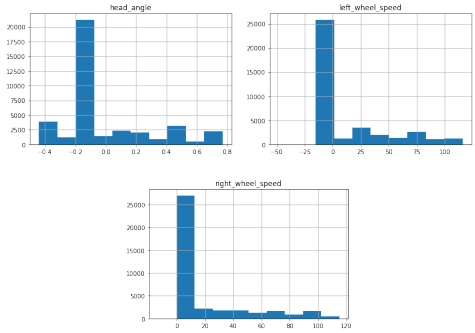

Figure 6: (Cozmo) Top Left: head angle ranges, Top Right: left wheel speed ranges, Bottom: right wheel speed ranges

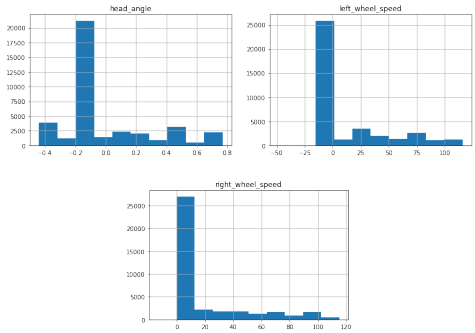

Figure 7: (Cozmo) Top Left: head angle ranges, Top Right: left wheel speed ranges, Bottom: right wheel speed ranges

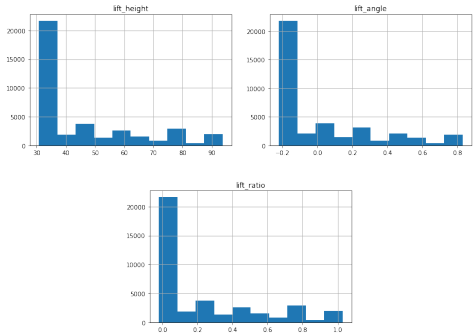

Figure 8: (Cozmo) Top Left: lift height ranges, Top Right: lift angle ranges, Bottom: lift ratio ranges

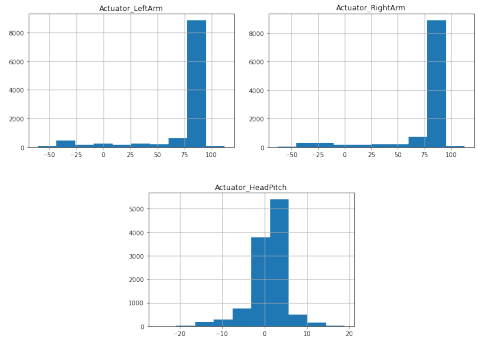

Figure 9: (Misty) Top Left: left arm ranges, Top Right: right arm ranges, Bottom: head pitch ranges

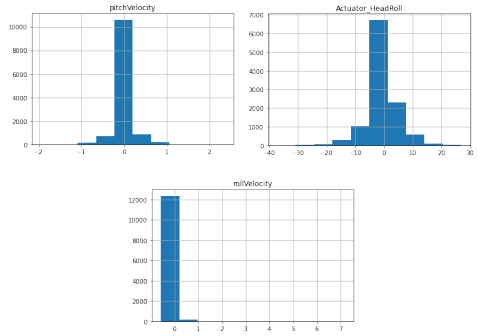

Figure 10: (Misty) Top Left: head pitch velocity ranges, Top Right: head roll ranges, Bottom: head roll velocity ranges

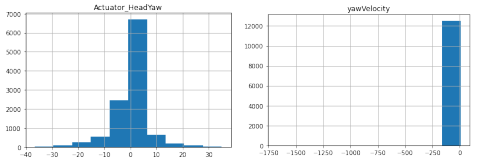

Figure 11: (Misty) Left: head yaw ranges, Right: head head yaw velocity ranges