Distinguishing Engagement Facets: An Essential Component for AI-based Healthcare

Hanan Salam

Received: date / Accepted: date

Abstract Engagement in Human-Machine Interaction is the process by which entities participating in the interaction establish, maintain, and end their perceived connection. It is essential to monitor the engagement state of patients in various AI-based healthcare paradigms. This includes medical conditions that alter social behavior such as Autism Spectrum Disorder (ASD) or Attention-Deficit/Hyperactivity Disorder (ADHD). Engagement is a multifaceted construct which is composed of behavioral, emotional, and mental components. Previous research has neglected the multi-faceted nature of engagement. In this paper, a system is presented to distinguish these facets using contextual and relational features. This can facilitate further fine-grained analysis. Several machine learning classifiers including traditional and deep learning models are compared for this task. A highest accuracy of 74.57% with an F-Score and mean absolute error of 0.74 and 0.23 respectively was obtained on a balanced dataset of 22242 instances with neural network-based classification.

Keywords Engagement Recognition · Healthcare · Affective Computing · Human-Robot Interaction

1 Introduction

During the last decade, researchers have demonstrated interest in enhancing the capabilities of robots to assist humans in their daily life. This requires incorporation of social intelligence within the robots which involves understanding different states of engagement. Research in Human-Machine Interaction (HMI) has depicted that engagement is a multi-faceted construct and consists of different components. It is very much important to be able to distinguish the facets before performing a deeper analysis. Corrigan et al. \cite{12} demonstrated that engagement is mainly composed of cognitive and affective components which are manifested by attention and enjoyment. According to O’brien et al. \cite{38}, engagement is characterized by
features like challenge, positive affect, endurability, aesthetic and sensory appeal, attention, feedback, variety/novelty, interactivity, and perceived user control. In the context of youth engagement in activities, Ramey et al. [12] proposed a model of psychological engagement having three components: cognitive like thinking or concentrating, affective like enjoyment, and relational like through connectedness to something. These suggest that when attempting automatic inference of user’s engagement state, it is important to consider this multi-faceted nature.

Application areas of Assistive Robotics include elderly care [9], helping people with medical conditions that alter social behavior such as children suffering from Autism Spectrum Disorder (ASD) [20] or people suffering from Adult Deficient Hyperactivity Disorder (ADHD) [23], coaching and tutoring [24, 52]. Fasola and Mataric [19] presented a Socially Assistive Robot (SAR) system designed to engage elderly users in physical exercise. Different variants of the robot’s verbal instructions were used to minimize the robot’s perceived verbal repetitiveness, and thus maintain the users’ engagement. Previous engagement detection approaches revolve around a binary classification-based approach (engaged vs. not engaged) [21, 32] or a multi-class approach (engagement level) [11, 41]. However, the multi-faceted nature is seldom considered.

In this paper, a framework that takes into account the multi-faceted nature of engagement is proposed. Such analysis allows to inform the implementation of fine-grained strategies based on a deeper understanding of user’s states. Engagement is modeled with an artificial neural network in terms of a spectrum of engagement states: mental, behavioural and emotional. We present a preliminary evaluation of this approach on an offline multi-party HRI corpus.

2 Related Work

Engagement in Human-Robot Interaction is defined as the process by which two (or more) participants establish, maintain, and end their perceived connection [45, 46]. Andrist et al. [1] analyzed an HRI dataset in terms of interaction type, quality, problem types, and the system’s failure points causing problems. Failure in the engagement component was found to be among the major identified problems cause during the interaction. This confirms that a highly performing engagement model is essential for the success of any HRI scenario [1].

Bohus & Horvitz [7] pioneered research on engagement in multi-party interaction. They explored disparate engagement strategies to allow robots to engage simultaneously with multiple users. There are multifarious studies based on multi-party interactions. Oertel et al. [39] studied in both individual and group level about the relationship amidst the participants’ gaze and speech behavior. Leite et al. [32] experimented with the generalization capacity of an engagement model. It was trained and tested on single-party and multi-party scenarios respectively. The opposite scenario was also considered. Salam et al. [44] conducted a study on engagement recognition in a triadic HRI scenario and showed that it is possible to infer a participant’s engagement state based on the other participants’ cues.

Most of engagement inference approaches revolved around identification of a person’s intention to engage. There has also been studies to detect whether the person is engaged/disengaged. Benkaouar et al. [5] presented a system to detect
disparate engagement phases. This includes intention to engage, engaged and disengaged. Foster et al. [21] attempted to detect whether a person intends to engage which is a bi-class problem. Leite et al. [32] attempted to identify disengagement in both group and individual interactions. Benet al. [4] also presented a system dedicated to the similar cause.

There are different works which focused on detecting different levels of engagement of a user. Michalowski et al. [36] distinguished different levels of engagement in the thick of present, interacting, engaged, and just attending. A system to distinguish two classes of engagement namely medium-high to high and medium-high to low engagement was presented by [11]. Bednarik et al. [3] distinguished disparate states of conversational engagement states. It includes no interest, following, responding, conversing, influencing and managing. They also modelled a bi-class problem having low/high conversational level. Oetrel et al. [39] distinguished 4 classes for group involvement namely high, low, leader, steering the conversation, and group is forming itself. Two models were developed in [17] focusing on not-engaged/engaged and not-engaged/normally-engaged/very-engaged state distinction. Frank et al. [22] differentiated 6 different states of engagement in the thick of disengagement, involved engagement, relaxed engagement, intention to act, action, and involved action.

Recently, [16] stated that engagement in HRI should be multi-faceted. Formulating binary/multi-class problems (engaged vs. not engaged) or a multi-class problem (engagement level) over this ignore the multi-faceted nature of engagement. Taking this multi-faceted nature into consideration is very important for the design of intelligent social agents. For instance, this can influence the implemented engagement strategies within the agent’s architecture. Some studies attempted to implement different strategies related to task and social engagement. For instance, [18] implemented a task engagement strategy which focuses on the task at hand and having users meta-cognitively reflect on the robot’s performance and a social engagement strategy which focuses on their enjoyment and having them meta-cognitively reflect on their emotions with respect to the activity and the group interactions.

Different features have been used to distinguish engagement states. Some of such features include contextual [31, 11, 44] attentional [51, 40], affective [11, 21, 17, 35] to name a few. Salam et al. [43] used person to detect both individual and group engagement. [26, 27] combined different aspects like backchannels, eye gaze, head nodding-based features to detect engagement level. Ben et al. [4] combined several attributes like speech and facial expressions, gaze and head motion, distance to robot to identify disengagement. Masui et al. [35] worked with facial Action Units and physiological responses.

Recent approaches explored deep learning architectures for the detection of engagement. Dewan et al. [17] used person-independent edge features and Kernel Principal Component Analysis (KPCA) within a deep learning framework to detect online learners’ engagement using facial expressions. [14] used CNN and LSTM networks to predict engagement level.

Contextual information is being used in social signal processing for quite some time. Kapoor et al. [30] combined context features in the form of game state with facial and posture features in an online educative scenario. Martinez and Yannakakis [33] used sequence mining for the prediction of computer game player affective states. Castellano et al. [11] explored task and social-based contextual
features. In another instance, the authors [10] used same contextual features for distinguishing interaction quality.

Relational feature have proven to be useful in multifarious instances. Curhan et al. [13] used dyad-based cues for predicting negotiation outcomes. Jayagopi et al. [28] adhered to group-based cues to understand typical behavior in small groups. Nguyen et al. [37] extracted relational audio-visual cues to detect the suitability of an applicant in a job interview. The features included audio and visual back-channeling, nodding while speaking, mutual short utterances and nods. It also includes [6] that used “looking-while-speaking” feature to understand personality impressions from conversational logs extracted from YouTube.

So far, context has been insufficiently investigated in the avenue of affective and cognitive states. Devillers et al. [15] highlights the importance of context in the assessment of engagement. They identified paralinguistic, linguistic, non-verbal, interactional, and specific emotional and mental state-based features as very important for engagement prediction. In this work, we investigate relational and contextual features for the recognition of a spectrum of engagement states. The features have been used in isolation as well as in combination to assess their engagement state distinction capability. These features have not been combined previously for detecting engagement facets. Compared to previous works, the proposed features model interaction context, the robot’s behavior and the behavioral relation between the participant in question and the other entities of the interaction.

3 Need of Engagement Recognition in Healthcare

Technological advancements have propagated to every field. There has always been efforts to automate tasks. Healthcare is one of the primal needs for society and it also has been touched by technology [48]. Several systems have come up to aid in automated healthcare and the well-being of people with medical conditions. In [23] a SAR is proposed, whose aim is to help children with ADHD to improve their educational outcome through social interaction with a robot. Another educational SAR was presented by [25]. This was targeted towards providing assistance in personalizing education in classrooms. The aim of the system is to provide assistance in personalizing education in classrooms. Children with Autism Spectrum Disorder are a target population for such personalized teaching systems. However, the systems do not include a user’s engagement analysis module. Such Socially Assistive systems can largely benefit from a fine-grained analysis of engagement. This will make the systems more human-like. There has been interest in automated screening and consultation to detect problems of the body and mind at an early stage. This can also help to reduce the initial load on doctors. It is very important for the patients to feel that they are interacting with their peers rather than a machine. The systems need to process both audio and visual cues in order to properly understand patients. While the patients are interacting with the automated systems, several states of engagement needs to be monitored simultaneously. This includes level of concentration, different reactions, spontaneity to name a few. Such states of engagement portray useful information about a patient’s health. These engagement states can be categorized into a broader spectrum of behavioral, mental and emotional states. Distinguishing the engagement facet is important at the outset for a deeper analysis. This can pave the way for systems
which would better understand the condition of patients by reading their body language and not merely match spoken symptoms. This will especially be useful in treating and understanding mental conditions where the body language is a vital aspect. In the case of psychological problems, patients are often engaged into conversations regarding disparate aspects by doctors wherein the patient’s body language serves a vital pointer towards the mental condition.

4 Methodology

The proposed framework is composed of 3 steps. First, a multi-party HRI corpus is annotated in terms of engagement facets. Then, different contextual and relational features are extracted. Finally, a neural network based classification model is used to classify the different engagement facets. Fig. 1 presents an illustration of the proposed methodology.

4.1 Data Corpus

In this section, the data corpus along with the disparate engagement annotations is discussed.

4.2 Video Material

We use 4 interactions of 8 participants from the conversational HRI data corpus ‘Vernissage’ [29]. It is a multi-party interaction amidst the humanoid robot NAO and 2 participants. The interaction has different contexts which can mainly be differentiated into 2 parts. The 1st is where the robot describes several paintings hanged on a wall. In the 2nd the robot performs a quiz with the volunteers related to art and culture. This was done in order to encompass different variations for the engagement states.

The average length per interaction is nearly 11 minutes. NAO’s internal camera was used to record the clips. This provided the front view. 3 other cameras were used.

Fig. 1: Proposed methodology
Fig. 2: Organisation of the recording room. NAO (orange), participants typical positions (gray circles), cameras (HD: red, VICON: blue), wizard feedback (green), paintings (green lines), windows (blue lines), VICON coordinate system (red), head pose calibration position (P1 and P2).

The corpus has annotations for non-verbal behaviors of the participants. It also contains robot’s speech and action in the log file of the robot.

4.3 Engagement Annotations

Engagement labels were assigned to 3 categories namely mental, behavioral and emotional. These were annotated when the participants manifested one of the following states: thinking, listening, positive/negative reaction, responding, waiting for feedback, concentrating, and listening to the other participant. The annotations were performed by 2 people with the aid of Elan annotation tool. They watched every video 2 times (once with the perspective of 1 participant). Discrete segments were annotated and it was stopped as soon as a change was observed. The Mean inter-rater Cronbach’ Alpha coefficient was 0.93. This points to the reliability of the annotations. The details of each category is as follows.

**Mental states** – A segment was assigned mental state label when the participant manifested one of the following mental states:

- **Listening (EL)**: The participant is listening to NAO;
- **WaitingFeedback (EWF)**: The participant is waiting for NAO’s feedback after he/she had answered a question;
- **Thinking (ETh)**: The participant is thinking about the response to a question asked by NAO;
- **Concentrating (EC)**: The participant is concentrating with NAO;
- **ListeningPerson2 (ELP2)**: The participant is listening to the other who is answering NAO.

2 https://tla.mpi.nl/tools/tla-tools/elan/
Table 1: Details on the number of annotated instances in each class

| State    | Number of instances |
|----------|---------------------|
| Behavioral | 10331              |
| Emotional | 7414               |
| Mental    | 80902              |
| Total     | 98647              |

**Behavioral states** – A segment was assigned a behavioral state label when the participant manifested the following behavioral state:

– *Responding (ER)*: The participant is responding to NAO;

**Emotional states** – A segment was assigned an emotional state label when the participant manifested one of the following emotional states:

– *PositiveReaction (EPR)*: The participant shows a positive reaction to NAO.
– *NegativeReaction (ENR)*: The participant shows a negative reaction to NAO.

The details regarding the number of annotated instances for each class is presented in Table 1.

4.4 Extracted Features

In this study, we used the annotated cues from Vernissage corpus. Moreover, we extracted additional metrics which were computed from the existing ones. They were categorized into two categories: 1) contextual and 2) relational.

Contextual features deal with either the different entities of an interaction like the robot utterance, addressee and topic of speech or behavioral aspects of the participant that concern the interaction context like visual focus of attention and addressee.

Relational features encode the behavioral relation between the participants and the robot. Fig. 3 illustrates the features groups used in our study.

4.4.1 Contextual Features

Interaction amidst entities involves both entities and connection. While inferring the engagement state of an interacting person, we consider behavior of the person as well as our behavior. Thus, an automated engagement identification system should also consider the same.

Consequently, we employ different contextual features that describes the participant’s behavior with respect to the other entities. Moreover, for a dialogue of the robot, we extract the robot’s utterance, addressee and topic of speech.

**Participant:**

1) **Visual Focus Of Attention (VFOA):** Gaze in human-human social interactions is considered as the primary cue of attention [34, 47]. We use VFOA ground truth of every participant which were annotated with 9 labels.
2) **VFOA Shifts**: Gaze shifts indicate people’s engagement/ disengagement with specific environmental stimuli [2]. We define VFOA shift as the moment when a participant shifts attention to a different subject. This feature is binary and is computed from the VFOA labels.

3) **Addressee**: When addressing somebody, we are engaged with him/her. Similarly, in the context of HRI, when a participant addresses someone other than the robot, he/she is disengaged from the robot. Adressee annotations used from the corpus and are annotated into 6 Classes: {NoLabel, Nao, Group, PRight, PLeft, Silence}.

**Robot**: Starting from the robot’s conversation logs, the following were extracted.

1) **Utterances**: The labels {Speech, Silence} were assigned to frames depending on the robot’s speech activity.

2) **Addressee**: The addressee of the robot was detected using predefined words from its speech. The following labels were assigned {Person1, Person2, Group-Explicit, GroupPerson1, GroupPerson2, Person1Group, Person2Group, Group, Silence}. ‘GroupExplicit’ label refers to such segments where the robot was explicitly addressing both participants. ‘GroupPersonX’ /X ∈ (1, 2) corresponds to segments where the robot addresses the group then ‘PersonX’ while ‘PersonXGroup’ represents the inverse.

3) **Topic of Speech**: This was identified using a keyword set. These were related to disparate paintings available in the scene {manray, warhol, arp, paintings}. Frames were allotted labels based on them.

### 4.4.2 Relational Features

We extract a set of Relational Features describing robot’s and participants’ behaviors synchrony and alignment. These include, among others, mutual gaze and...
laughter. A logical AND operation was used between participants’ and robot’s features time series for obtaining mutual events occurrence. Fig. 4 shows an example of participants’ mutual laughter extraction.

![Fig. 4: Example of relational cues extraction. This corresponds to participants’ mutual laughter detection using logical AND from laughter time series.](image)

**Participant-Robot Features:**

1) **Gaze-Speech Alignment:** We extracted events where a participant looks at objects corresponding to the robot’s topic of speech. This indicates that the participant is listening to the robot and is interested in what it is saying.

2) **P1 talks to P2/Robot Speaks:** This refers to events where the participants speak with each other during the robot’s speech. This may signal a disengagement behavior.

**Person1-Person2 Features:**

1) **Participants Mutual Looks:** This refers to events where the participants look at each other. Though this may signal disengagement but it may also signal engagement as it might be a reaction to the robot’s speech.

2) **Participants Mutual Laughter:** This refers to events where the two participants laugh together. This represents reaction to the robot’s speech.

3) **P1 Looks at P2/ P2 Talks to Robot:** This represents events where the passive participant looks at active participant while he/she is talking to the robot. Though this may appear to be disengagement, but analysis revealed the inverse.

The total number of features is 39. There were 34 contextual features and 5 relational features.

### 4.5 Classification: Artificial Neural Network

A multi layer perceptron is a feed forward neural network. It is mainly composed of 3 layers namely input layer, hidden layer and output layer. Each of the layers consist of neurons which are connected to the next layer. The number of neurons in the input layer corresponds to the number of feature inputs while that of the output layer are equal to the number of classes. The connections are associated with certain weights which are tuned at the time of training. If a neuron $N$ has connection from 3 other neurons ($n_1$, $n_2$, and $n_3$) of the previous layer with weights as $W_1$, $W_2$, and $W_3$, then the net input $N_I$ to $N$ can be represented as:

$$N_I = n_1 \times W_1 + n_2 \times W_2 + n_3 \times W_3$$ (1)
Table 2: Performance of different features

| Feature              | Accuracy (%) |
|----------------------|--------------|
| Relational           | 82.51        |
| Contextual           | 87.48        |
| Relational+Contextual| 87.77        |

Table 3: Confusion matrix for the combined feature set

|       | Behavioral | Emotional | Mental  |
|-------|------------|-----------|---------|
| Behavioral | 4955       | 134       | 5242    |
| Emotional  | 269        | 1966      | 5179    |
| Mental     | 871        | 373       | 79658   |

The output of a neuron is controlled by an activation function for instance a \( \tanh \) function. Thus the output \( N_O \) of neuron \( N \) can be represented as:

\[
N_O = \tanh(N_I)
\]

The output of a neural network is a set of probabilities. Each of the probabilities correspond to belongingness of an instance to a particular class. The class with the highest probability is assigned as the label of a particular instance. Initially, the network was trained with 500 iterations. The learning rate was fixed at 0.3 while the momentum was chosen to be 0.2. The number of neurons was 3 for all the feature sets. In the case of the input layer, the combined feature set had 39 neurons. Similarly there were 34 and 5 neurons in the input layer for the contextual and relational feature sets respectively.

5 Experiments and Results

The proposed approach is evaluated using 5-fold cross validation scheme to avoid possible bias. This ensured that every instance was subjected to training and testing at least once.

5.1 Performance Analysis of Features

The contextual, relational and a combination of both was fed to the neural network whose results are summarized in Table 2. It is seen that the contextual features performed better as compared to the relational features. The combined features performed best among the 3. Its performance was slightly better than the Contextual features (0.29% more). The confusion matrix for this setup is presented in Table 3. It is seen that most of the confusions occurred with the class Mental. One probable reason for this is the significantly large number of instances for this class as compared to the rest. Further tuning of the neural network was done with the combined features due to its superior performance over the rest.
Table 4: Performance for different learning rates on the combined feature set

| Learning Rate | 0.1   | 0.2   | 0.3   | 0.4   | 0.5   |
|---------------|-------|-------|-------|-------|-------|
| Accuracy (%)  | 87.80 | 87.79 | 87.77 | 87.77 | 87.78 |

Table 5: Performance for learning rate=0.1 on the combined feature set

|         | Behavioral | Emotional | Mental |
|---------|------------|-----------|--------|
| Behavioral | 4989       | 151       | 5191   |
| Emotional | 271        | 2045      | 5098   |
| Mental   | 882        | 438       | 79582  |

Table 6: Performance for different momentum values on the combined feature set

| Momentum | 0.1   | 0.2   | 0.3   | 0.4   | 0.5   |
|----------|-------|-------|-------|-------|-------|
| Accuracy (%) | 87.80 | 87.77 | 87.80 | 87.77 | 87.72 |

Table 7: Confusion matrix for momentum=0.1 on combined feature set

|         | Behavioral | Emotional | Mental |
|---------|------------|-----------|--------|
| Behavioral | 5016       | 142       | 5173   |
| Emotional | 280        | 1972      | 5162   |
| Mental   | 904        | 375       | 79623  |

5.2 Performance Analysis w.r.t Learning Rate

Next, the learning rate was varied from 0.1 to 0.5 with a step of 0.1 whose results are shown in Table 4. It is seen that the best result was obtained for a learning rate of 0.1. The performance improved over the initial learning rate of 0.3. The performance dropped on increasing the learning rate. The confusion matrix for this setup is presented in Table 5. The individual recognition rates increased for both behavioral and emotional class as compared to the initial setup. However, the recognition rate dropped slightly for the mental class.

5.3 Performance Analysis w.r.t Momentum Values

Next, the performance was analyzed for different momentum values from 0.1 to 0.5 with a step of 0.1. The obtained results are listed in Table 6. The best result was obtained for a momentum value of 0.1 which reduced on increasing the momentum. The confusion matrix for this setup is presented in Table 7.

5.4 Performance Analysis w.r.t Training Iterations

The training iterations were varied from 100-600 with a step of 100. The best performance was obtained for both 400 and 500 iterations. However, 400 iterations was considered because a system which is trainable with lower iterations is always beneficial. The results along with the associated errors is presented in Table 8.
Table 8: Performance for different training iterations along with the associated error.

| Iterations | 100  | 200  | 300  | 400  | 500  | 600  |
|------------|------|------|------|------|------|------|
| Accuracy (%) | 87.76 | 87.76 | 87.75 | 87.77 | 87.77 | 87.76 |
| Mean Absolute Error | 0.1262 | 0.1258 | 0.1258 | 0.1258 | 0.1259 | 0.1259 |

Table 9: Consolidated value of the hyperparameters.

| Hyperparameter | Value |
|----------------|-------|
| Learning rate  | 0.1   |
| Momentum       | 0.1   |
| Training epoch | 400   |

Table 10: Confusion matrix for best performance Momentum=0.1, Learning Rate=0.1, Training iterations=400

|         | Behavioral | Emotional | Mental  |
|---------|------------|-----------|---------|
| Behavioral | 5022       | 154       | 5155    |
| Emotional  | 264        | 2039      | 5111    |
| Mental     | 903        | 431       | 79568   |

5.5 Performance Using Tuned Parameters

The best momentum, learning rate, and training epoch values is presented in Table 9. These values were used in conjunction which produced an accuracy of 87.82%. This was the overall highest in the experiments. The confusion matrix for this setup is presented in Table 10. The performance increased by 0.06% as compared to the initial setup.

It is seen that almost 1.49% of the behavioral type instances were confused to be emotional while almost 3.56% of the emotional confused to be that of behavioral. A major chunk of both of them were confused to be of mental type. This is possibly due to the unbalanced data. The precision values for the behavioral, mental, and emotional class types were 0.81, 0.7, and 0.89 respectively.

5.6 Performance For Balanced Dataset

As the data was highly imbalanced, so a subset of the data was drawn having equal number of instances per class totalling to 22242 instances. The combined features (contextual+relational) for this dataset was fed to the network with the best performing hyperparameters. The confusion matrix is presented in Table 11. It is noted that the best performance was obtained for behavioral class which produced an accuracy of 85.6%. This was followed by mental where an accuracy of 72.1% was obtained. The lowest performance was obtained for emotional class (Accuracy=66%). The highest confused pair was emotional-mental which led to almost 10.59% of the total instances to be misclassified. The mean absolute error was 0.23. The values for different performance metrics for each of the classes is presented in Table 12.
Table 11: Confusion matrix for balanced dataset

|       | Behavioral | Emotional | Mental |
|-------|------------|-----------|--------|
| Behavioral | 6347       | 384       | 683    |
| Emotional  | 1318       | 4893      | 1203   |
| Mental     | 916        | 1153      | 5345   |

Table 12: Class-wise values for performance metrics on the balanced dataset.

| Metrics                  | Behavioral | Emotional | Mental |
|--------------------------|------------|-----------|--------|
| True positive rate       | 0.856      | 0.660     | 0.721  |
| False positive rate      | 0.151      | 0.104     | 0.127  |
| Precision                | 0.740      | 0.761     | 0.739  |
| Recall                   | 0.856      | 0.660     | 0.721  |
| F-score                  | 0.794      | 0.707     | 0.730  |

Table 13: Comparative analysis of the performance of standard classifiers on unbalanced and balanced dataset.

| Classifier               | Accuracy (%) |
|--------------------------|--------------|
|                          | Unbalanced   | Balanced    |
| BayesNet                 | 85.96        | 69.61       |
| Naive Bayes              | 84.41        | 68.61       |
| Simple Logistic          | 86.74        | 70.35       |
| SVM                      | 86.72        | 69.68       |
| RBF Network              | 85.17        | 65.13       |
| Recurrent Neural Network | 84.29        | 57.68       |
| Artificial Neural Network| 87.82        | 74.57       |

5.7 Comparative Study

The performance of different classifiers was analyzed. The combination feature was used in this case due to its better performance. The compared classifiers consist of BayesNet, Naive Bayes, Simple logistic, SVM, RBF network, and Recurrent neural network. However, none of the classifiers produced better results than the neural network. Simple Logistic produced the best result after neural network. It was only 1.08% lower than the best result. The performance of SVM was also very close to simple logistic. It was only 0.2% less than Simple Logistic. The least accuracy of 84.29% was obtained for Recurrent neural network. A slight improvement of 0.12% was obtained for Naive Bayes.

It is observed that the performance of the balanced set lacked behind the unbalanced dataset. However, the results were not biased by the mental class. As the number of instances were extremely large for mental, the trained model was possibly biased towards mental class which influenced the final accuracy of the system. In this case, also a similar trend in the ranks of the classifiers was observed as compared to the unbalanced data. However, the performance of Recurrent neural network was significantly low as compared to that of the neural network. The second best performer, (simple logistic) was also approximately 5.7% lower than the highest result.
6 Conclusions and future work

In this paper, we proposed a system to detect different facets of engagement states: mental, emotional, and behavioral. This is essential for a deeper analysis of the user’s engagement by machines. Experiments were performed with contextual and relational features and their combination. The best result was obtained by combining the features. Experiments were performed on both unbalanced and balanced datasets and a highest F-score of 0.74 was obtained for the balanced dataset. In future, we plan to validate the framework on a larger dataset. The dataset will be expanded both in terms of the participants and the observations. We plan to work with individual features to improve the system’s performance and perform a deep grained analysis of the different states. We will also explore deep learning-based approaches and unsupervised approaches towards detection of engagement state types and thereafter finer classification. Deep learning will be used not only for data classification but also for feature extraction. In real world, data is mostly available in the form of streams which motivates us to explore active learning approaches as well.

Conflict of Interest

Authors declare no conflict of interest.

References

1. Andrist, S., Bohus, D., Kamar, E., Horvitz, E.: What went wrong and why? diagnosing situated interaction failures in the wild. In: International Conference on Social Robotics, pp. 293–303. Springer (2017)
2. Baron-Cohen, S.: Mindblindness: An essay on autism and theory of mind. MIT press (1997)
3. Bednarik, R., Eivazi, S., Hradis, M.: Gaze and conversational engagement in multiparty video conversation: An annotation scheme and classification of high and low levels of engagement. In: Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction, pp. 1–6. ACM (2012)
4. Ben-Youssef, A., Varni, G., Essid, S., Clavel, C.: On-the-fly detection of user engagement decrease in spontaneous human–robot interaction using recurrent and deep neural networks. International Journal of Social Robotics 11(5), 815–828 (2019)
5. Benkaouar, W., Vaufreidy, D.: Multi-sensors engagement detection with a robot companion in a home environment. In: Workshop on Assistance and Service robotics in a human environment at IEEE International Conference on Intelligent Robots and Systems (IROS), pp. 45–52 (2012)
6. Biel, J., Gatica-Perez, D.: The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. Multimedia, IEEE Transactions on 15(1), 41–55 (2013)
7. Bohus, D., Horvitz, E.: Models for multiparty engagement in open-world dialog. In: Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 225–234. Association for Computational Linguistics (2009)
8. Bouguelia, M.R., Nowaczyk, S., Santosh, K., Verikas, A.: Agreeing to disagree: active learning with noisy labels without crowdsourcing. International Journal of Machine Learning and Cybernetics 9(8), 1307–1319 (2018)
9. Broadbent, E., Jayawardena, C., Kerse, N., Stafford, R., MacDonald, B.: Human-robot interaction research to improve quality of life in elder care—an approach and issues. In: 25th Conference on Artificial Intelligence (AAAI). Workshop on Human-Robot Interaction in Elder Care, pp. 13–19 (2011)
10. Castellano, G., Leite, I., Paiva, A.: Detecting perceived quality of interaction with a robot using contextual features. Autonomous Robots pp. 1–17 (2016)
11. Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A., McOwan, P.W.: Detecting engagement in hri: An exploration of social and task-based context. In: International Conference on Privacy, Security, Risk and Trust and International Conference on Social Computing, pp. 421–428. IEEE (2012)
12. Corrigan, L.J., Peters, C., Küster, D., Castellano, G.: Toward Robotic Socially Believable Behaving Systems - Volume I : Modeling Emotions, chap. Engagement Perception and Generation for Social Robots and Virtual Agents, pp. 29–51. Springer International Publishing (2016)
13. Curhan, J.R., Pentland, A.: Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. Journal of Applied Psychology 92(3), 802 (2007)
14. Del Dauchetto, F., Baxter, P., Hanheide, M.: Are you still with me? continuous engagement assessment from a robot’s point of view. arXiv preprint arXiv:2001.03515 (2020)
15. Devillers, L., Duplessis, G.D.: Toward a context-based approach to assess engagement in human-robot social interaction. In: Dialogues with Social Robots, pp. 293–301. Springer (2017)
16. Devillers, L., Rosset, S., Duplessis, G.D., Bechade, L., Yemez, Y., Turkar, B.B., Sezgin, M., Erzin, E., El Haddad, K., Dupont, S., et al.: Multifaceted engagement in social interaction with a machine: The joker project. In: 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 697–701. IEEE (2018)
17. Dewan, M.A.A., Lin, F., Wen, D., Mushed, M., Uddin, Z.: A deep learning approach to detecting engagement of online learners. In: IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, pp. 1895–1902. IEEE (2018)
18. El Hamamsy, L., Johal, W., Asselborn, T., Naïr, J., Dillenbourg, P.: Learning by collaborative teaching: An engaging multi-party co-writer activity. In: 28th IEEE International Conference on Robot and Human Interactive Communication (ROMAN), pp. 1–8. IEEE (2019)
19. Fasola, J., Mataric, M.: A socially assistive robot exercise coach for the elderly. Journal of Human-Robot Interaction 2(2), 3–32 (2013)
20. Feil-seifer, D., Viterbi, U., et al.: Development of socially assistive robots for children with autism spectrum disorders. Tech. rep., Center for Robotics and Embedded Systems (2009)
21. Foster, M.E., Gaschler, A., Giuliani, M.: How can i help you: comparing engagement classification strategies for a robot bartender. In: Proceedings of the 15th ACM on International conference on multimodal interaction, pp. 255–262. ACM (2013)
22. Frank, M., Tofbghi, G., Gu, H., Fruchter, R.: Engagement detection in meetings. arXiv preprint arXiv:1608.08711 (2016)
23. Fridin, M., Yaakobi, Y.: Educational robot for children with adhd/add. In: Architectural Design, International Conference on Computational Vision and Robotics, Blubanewar, INDIA, pp. 1–7 (2011)
24. Greczek, J., Mataric, M.: Expanding the computational model of graded cueing: Robots encouraging health behavior change. In: 29th AAAI Conference on Artificial Intelligence, pp. 1–2 (2014)
25. Greczek, J., Short, E., Clabaugh, C.E., Swift-Spong, K., Mataric, M.: Socially assistive robotics for personalized education for children. In: AAAI Fall Symposium Series, pp. 1–3 (2014)
26. Inoue, K., Lala, D., Takanashi, K., Kawahara, T.: Engagement recognition in spoken dialogue via neural network by aggregating different annotators’ models. In: Interspeech, pp. 616–620 (2018)
27. Inoue, K., Lala, D., Takanashi, K., Kawahara, T.: Latent character model for engagement recognition based on multimodal behaviors. In: 9th International Workshop on Spoken Dialogue System Technology, pp. 119–130. Springer (2019)
28. Jayagopi, D., Sanchez-Cortes, D., Otsuka, K., Yamato, J., Gatica-Perez, D.: Linking speaking and looking behavior patterns with group composition, perception, and performance. In: Proceedings of the 14th ACM international conference on Multimodal interaction, pp. 433–440. ACM (2012)
29. Jayagopi, D.B., Sheikh, S., Klotz, D., Wienke, J., Odobe, J.M., Wrede, S., Khalidov, V., Nguyen, L., Wrede, B., Gatica-Perez, D.: The vernissage corpus: A multimodal human-robot-interaction dataset. Tech. rep., Bielefeld University (2012)
30. Kapoor, A., Picard, R.W.: Multimodal affect recognition in learning environments. In: 13th annual ACM international conference on Multimedia, pp. 677–682. ACM (2005)
31. Kapoor, A., Picard, R.W., Ivanov, Y.: Probabilistic combination of multiple modalities to detect interest. In: Proceedings of the 17th International Conference on Pattern Recognition, vol. 3, pp. 969–972. IEEE (2004)
32. Leite, I., McCoy, M., Ulman, D., Salomons, N., Scassellati, B.: Comparing models of disengagement in individual and group interactions. In: Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction, pp. 99–105. ACM (2015)
33. Martínez, H.P., Yannakakis, G.N.: Mining multimodal sequential patterns: a case study on affect detection. In: Proceedings of the 13th international conference on multimodal interfaces, pp. 3–10. ACM (2011)
34. Mason, M.F., Tatik, E.P., Macrae, C.N.: The look of love gaze shifts and person perception. Psychological Science 16(3), 236–239 (2005)
35. Masui, K., Okada, G., Tsumura, N.: Measurement of advertisement effect based on multimodal emotional responses considering personality. ITE Transactions on Media Technology and Applications 8(1), 49–59 (2020)
36. Michalowski, M.P., Sabanovic, S., Simmons, R.: A spatial model of engagement for a social robot. In: 9th IEEE International Workshop on Advanced Motion Control, pp. 762–767. IEEE (2006)
37. Nguyen, L.S., Frauenthor, D., Mast, M.S., Gatica-Perez, D.: Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. IEEE Transactions on Multimedia 6(4), 1038–1051 (2013)
38. O’Brien, H.L., Toms, E.G.: What is user engagement? a conceptual framework for defining user engagement with technology. Journal of the American Society for Information Science and Technology 59(6), 938–955 (2008)
39. Oertel, C., Salvi, G.: A gaze-based method for relating group involvement to individual engagement in multiparty dialogue. In: Proceedings of the 15th ACM on International conference on multimodal interaction, pp. 99–106. ACM (2013)
40. Papadopoulos, F., Corrigan, L.J., Jones, A., Castellano, G.: Learner modelling and automatic engagement recognition with robotic tutors. Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACH 2013 pp. 730–744 (2013)
41. Peters, C., Asteriadi, S., Karpouzis, K.: Investigating shared attention with a virtual agent using a gaze-based interface. Journal on Multimodal User Interfaces 3(1-2), 119–130 (2010)
42. Ramey, H.L., Rose-Krasnor, L., Busseri, M.A., Gadbois, S., Bowker, A., Findlay, L.: Measuring psychological engagement in youth activity involvement. Journal of adolescence 45, 237–249 (2015)
43. Salam, H., Celiktutan, O., Hupont, I., Gunes, H., Chetouani, M.: Fully automatic analysis of engagement and its relationship to personality in human-robot interactions. IEEE Access 5, 705–721 (2016)
44. Salam, H., Chetouani, M.: Engagement detection based on multi-party cues for human robot interaction. In: International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 341–347. IEEE (2015)
45. Sidner, C.L., Kidd, C.D., Lee, C., Lesh, N.: Where to look: a study of human-robot engagement. In: 9th international conference on Intelligent user interfaces, pp. 78–84. ACM (2004)
46. Sidner, C.L., Lee, C., Kidd, C.D., Lesh, N., Rich, C.: Explorations in engagement for humans and robots. Artificial Intelligence 166(1), 140–164 (2005)
47. Sidner, C.L., Lee, C., Lesh, N.: Engagement when looking: behaviors for robots when collaborating with people. In: Diab: Proceedings of the 7th workshop on the Semantic and Pragmatics of Dialogue, pp. 123–130. University of Saarland (2003)
48. Thelissou, E., Sharma, K., Salam, H., Dignum, V.: The general data protection regulation: An opportunity for the hci community? In: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–8 (2018)
49. Vajda, S., Santosh, K.: A fast k-nearest neighbor classifier using unsupervised clustering. In: International conference on recent trends in image processing and pattern recognition, pp. 185–193. Springer (2016)
50. Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Slaetjes, H.: Elan: a professional framework for multimodality research. In: Proceedings of 5th the International Conference on Language Resources and Evaluation (LREC), pp. 1556–1559 (2006)
51. Yun, S.S., Choi, M.T., Kim, M., Song, J.B.: Intention reading from a fuzzy-based human engagement model and behavioural features. International Journal of Advanced Robotic Systems (2012)

52. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2879–2886. IEEE (2012)