Context Matters: Understanding Socially Appropriate Affective Responses Via Sentence Embeddings

Youssef Mohamed[1], Séverin Lemaignan[2], Arzu Güneysu[1,3], Patric Jensfelt[1] and Christian Smith[1]

[1] KTH: The Royal Institute of Technology
[2] PAL Robotics
[3] Umeå University

Abstract. As AI systems increasingly engage in social interactions, comprehending human social dynamics is crucial. Affect recognition enables systems to respond appropriately to emotional nuances in social situations. However, existing multimodal approaches lack accounting for the social appropriateness of detected emotions within their contexts.
This paper presents a novel methodology leveraging sentence embeddings to distinguish socially appropriate and inappropriate interactions for more context-aware AI systems. Our approach measures the semantic distance between facial expression descriptions and predefined reference points. We evaluate our method using a benchmark dataset and a real-world robot deployment in a library, combining GPT-4(V) for expression descriptions and ada-2 for sentence embeddings to detect socially inappropriate interactions.
Our results underscore the importance of considering contextual factors for effective social interaction understanding through context-aware affect recognition, contributing to the development of socially intelligent AI capable of interpreting and responding to human affect appropriately.

Keywords Social representation, embeddings, machine learning, human-robot interaction

1 Introduction

The ability to navigate social contexts and detect socially appropriate behaviors is a critical requirement for intelligent systems designed to interact seamlessly with humans. Social appropriateness ensures that expressions and actions align with societal norms and situational expectations, fostering natural and harmonious interactions. For instance, a joyful expression at a funeral would be considered socially inappropriate, while the same expression at a celebratory event would be deemed appropriate [26]. This highlights the dynamic and context-dependent nature of social appropriateness, which intelligent systems must adeptly navigate to maintain effective communication and foster positive human-agent relationships

[19,6,20]. Moreover, the concept of appropriateness can vary significantly across cultures and subcultures, adding another layer of complexity to the challenge of defining and modeling social norms in intelligent systems [11].

Despite significant advancements in leveraging multimodal signals, such as facial expressions, speech, and physiological signals, to infer affective states [30,19,18], capturing the broader contextual factors that govern social appropriateness remains a critical challenge. While these modalities provide valuable insights into understanding affective states, they often fall short in encapsulating the full spectrum of social cues and contextual nuances essential for determining what constitutes socially appropriate behavior in a given situation. Addressing this gap necessitates the incorporation of contextual information that can modulate the interpretation of affective signals, enabling the generation of socially appropriate responses.

Models like Word2Vec [15] and GloVe [22] have revolutionized the understanding and processing of language, enabling more nuanced interpretations of textual data. Recent breakthroughs in natural language processing (NLP) have introduced sentence embeddings—a transformative approach that represents sentences as vectors in a high-dimensional space, capturing their semantic relationships [16,22]. These embeddings have been effectively applied in various domains, including sarcasm detection [1], demographic feature extraction [28], and analyzing social group attitudes and stereotypes [5]. Their ability to capture complex linguistic and contextual information holds promise for discerning social appropriateness in human-agent interactions.

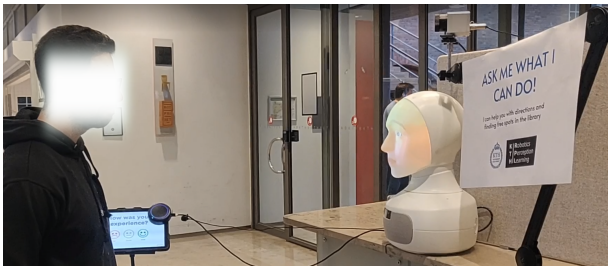


Fig. 1: The setup of the interaction. A Furhat robot is positioned in the library entrance to interact with visitors, who are encouraged to engage with it through a sign reading "Ask me what I can do!".

Our research aims to leverage the social knowledge embedded in language models to enable context-aware affect recognition [31,10,24,29]. The core idea is to capture the semantic nuances of facial expressions and contextual information through embeddings and assess the appropriateness of facial expressions within a given context by measuring the proximity between these embeddings.

Our first contribution is investigating to which extent sentence embeddings can capture contextual information from textual descriptions of facial expressions,

enabling the distinction between appropriate and inappropriate social interactions in neutral and negative contexts. We hypothesize that sentences describing facial expressions with similar social semantic meanings will be located closer to specific reference points in the embedding space. The distances between these points and the sentence embeddings are expected to correspond to appropriate and inappropriate interactions when the relevant context is provided.

Secondly, we explore whether the social capabilities exhibited by the embedding models can be applied to an in-the-wild dataset containing more natural facial expressions. We anticipate that by collecting a dataset in a real-world library context and representing affective language within an embedding space, F1 scores of the classification will be highest when using the correct context in which the library dataset was collected. Conversely, we expect these scores to decrease when an incorrect or opposite context is introduced.

2 Background and Related Work

In dynamic environments, robots must adapt continuously to maintain effective interactions. Dautenhahn [6] discusses the concept of socially intelligent robots, emphasizing the multifaceted nature of human-robot interaction. The paper highlights several key dimensions, including the robot's embodiment, its ability to express and perceive emotions, and its capacity for social learning. Particularly relevant to our work is Dautenhahn's emphasis on the importance of robots being able to recognize and interpret dynamic social cues. These cues can vary significantly based on context, cultural norms, and individual differences. The study underscores the need for robots to adjust their behavior in response to real-time changes in their social environment, suggesting that adaptive mechanisms are crucial for successful human-robot interaction.

Integrating and reasoning over contextual information remains a challenge. Traditional methods, like using image backgrounds as context, often involve complex computational processes [27,17]. For instance, Lee et al. [27] propose a context-aware emotion recognition network that requires hiding human faces in visual scenes and employing attention mechanisms to extract contextual information. Similarly, Mittal et al. [17] introduce a multimodal context-aware model that combines multiple human modalities with background semantic information and sociodynamic interactions, necessitating the use of self-attention-based CNNs and depth maps. While these approaches yield good results, they are often tailored to specific datasets and do not generalize to other scenarios that they were not trained on which is a major downside particularly in HRI applications, where robots must navigate diverse and dynamic social environments.

One promising approach is the use of sentence embeddings, which can encode both linguistic information and social interaction cues [12,2]. Advanced models such as BERT [8] and its derivatives, including RoBERTa [13] and GPT-3 [4], have demonstrated their capability to interpret contextually rich social interactions.

Lemaignan et al. [12] introduce the concept of social embeddings—compact mathematical representations of social situations derived from pre-trained large

language models. These embeddings are generated by creating textual descriptions of a robot's social environment and converting them into vector representations. The study details the methodology for constructing these embeddings, analyzes properties like social similarity measurement, and demonstrates their application in social robots.

Building on this foundation, we propose an approach that extends the use of embeddings to detect socially appropriate situations with enhanced flexibility. Our method distinguishes itself by integrating facial expressions and contextual information using a contrastive embedding technique. We employ sentence embeddings to represent both facial expression descriptions and context, while introducing contrasting reference points to differentiate appropriate from inappropriate behaviors. This approach offers significant advantages by eliminating the need for additional datasets or training for new social situations and providing greater adaptability to diverse social contexts without extensive model modifications. By focusing on the versatility of sentence embeddings, our method aims to overcome the rigidity of existing techniques. This flexibility is crucial for developing socially adept robots capable of navigating the complexities of human social environments.

2.1 Terminology

In this section, we introduce the definitions of some of the terminology used in this paper.

Socially Appropriate/Inappropriate: Socially appropriate interactions conform to societal norms and expectations within a given context, fostering positive social dynamics. Socially inappropriate interactions, on the other hand, deviate from these norms, potentially leading to discomfort or misunderstandings [3].

Contrasting Reference Points: Contrasting reference points are predefined semantic descriptions used to evaluate appropriateness by creating a scale in the embedding space between two contrasting sentences. For instance, "*This is a hostile interaction*" and "*This is a friendly interaction*" are considered two contrasting sentences [16].

Negative/Positive/Neutral Context: Negative contexts are scenarios where negative emotions or reactions are anticipated and socially appropriate. For example, at a funeral, feelings of sadness or other low-valence emotions are expected. In contrast, positive contexts are situations where positive emotions or reactions are suitable and foster positive social interactions, such as happiness at a celebration. Neutral contexts, on the other hand, are defined as those that do not inherently suggest or expect any specific emotional response [9].

Facial Expressions Descriptions: Facial expression descriptions are textual representations of observed facial cues and emotions, generated using advanced models like GPT-4(V).

3 Datasets

We evaluate the proposed approach on two datasets: 1) the MMI Dataset containing posed facial expressions [21], and 2) an in-the-wild dataset collected through naturalistic human-robot interactions.

3.1 MMI Dataset

The MMI Facial Expression Database [21] contains 348 videos featuring 51 participants exhibiting a diverse range of facial expressions elicited by different emotional stimuli labeled with the six basic emotions: happiness, sadness, fear, disgust, surprise and neutral expressions. Each video begins with a neutral expression, followed by a gradual transition to the target emotional expression. The database provides detailed annotations, including the onset and apex frames for each expression.

The dataset was selected due to its simple labelling scheme, comprising six basic emotions. This well-defined ground truth for emotions and when they occur facilitated a more focused evaluation of the embedding model's capability to differentiate emotional expressions across neutral and negative contexts.

3.2 In-The-Wild dataset in a Library setting

For our in-the-wild dataset, raw video data was collected in a library context and converted into a series of frames captured at a rate of 1 frame per second.

The video/audio data collection was conducted over a period of one week at the university library, where a Furhat robot was stationed in the foyer serving as an informational assistant to visitors (see Figure 1). The robot was capable of recognizing and responding to basic queries about the library. However, the response mechanism was intentionally designed to lead to unsuccessful/less useful interactions. For example, if a visitor asked about *books*, the robot would respond *the books are on the shelves*.

The interaction started when the participant approached the Furhat robot, and was considered complete once the participant's head gaze shifted away from the robot.

This dataset was selected due to its well-defined environmental context and the robot's intentional programming to exhibit failures. This approach facilitated the creation of a balanced dataset, capturing the facial expressions of participants as they experienced these controlled robotic failures in a natural environment. By strategically inducing failures, we were able to elicit a diverse range of facial expressions.

Participants Data was gathered from a total of 90 interactions, amounting to 80 minutes of engagement with adults between 18 to 40 years old. However, 58 interactions were excluded from the analysis due to participants wearing face masks, missing audio, incomplete sessions, or noisy environments, resulting in 32

valid one-on-one interactions with the robot spanning 35 minutes. A new session started once the system no longer detected the face of the initial participant, assigning a new participant ID for the subsequent interaction.

Labeling The labelling process was meticulously designed to ensure the accurate identification of failures during the interactions between the robot and the user. Each interaction was evaluated by two independent labelers to determine the presence and duration of any interaction failures. Cohen's Kappa (κ) was calculated to quantify the agreement between annotators, yielding a κ value of 0.698, indicative of substantial inter-annotator agreement. The average duration identified for failures within these interactions was approximately 10 seconds. The outcomes of the labelling process were binary, with '0' indicating the absence of a failure and '1' signifying the presence of a failure.

The criteria provided to the labellers were explicit, focusing on two main aspects of the interaction:

- **Social Appropriateness:** A failure is marked if the robot's response falls outside the socially appropriate bounds. This includes responses that are contextually irrelevant or significantly misaligned with the expected conversational norms.
- **Speech Recognition Accuracy:** A failure is also identified if the robot's response indicates a misunderstanding or misinterpretation of the user's spoken input. This encompasses scenarios where the robot's answer is unrelated to the user's request due to errors in speech recognition.

Labellers were instructed to evaluate each interaction, marking the start and end times of any segment that met the failure criteria outlined above. This structured approach ensured a comprehensive assessment of the robot's performance in real-world interactions, focusing on critical aspects of social appropriateness and speech recognition accuracy. Instances labelled as failure cases were those where both annotators independently agreed on the presence of a failure. Any instances where the annotators disagreed were assigned a label of "0", indicating no failure present.

It's important to note that while our dataset labels indicate inappropriate reactions of the robot, we make the assumption that participants' facial expressions will correspond to these robot actions. This assumption is based on previous work by [25], which demonstrated the effectiveness of using human facial expressions to detect robot failures in human-robot interactions.

4 Methodology

Our methodology leverages sentence embeddings to detect socially appropriate situations in human-robot interactions. We propose a novel approach that utilizes pre-trained embedding models to capture the nuanced social knowledge inherent in language. The core of our method involves the following.

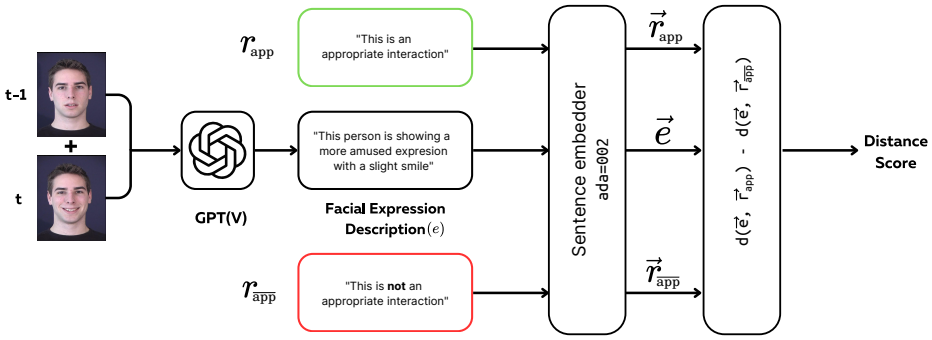


Fig. 2: The overall process for calculating the social appropriateness score of a facial expression within a given context. The system takes as input two images at time t and $t-1$, which are then processed by a GPT model to generate a facial expression description e. Additionally, two predefined reference points, r_{app} and r_{inapp}, representing socially appropriate and inappropriate interactions, respectively, are provided. An embeddings model is employed to transform the facial expression description e and the reference points r_{app} and r_{inapp} into their corresponding vector representations $\vec{e}$, $\vec{r}_{\text{app}}$, and $\vec{r}_{\text{inapp}}$. The social appropriateness score is then calculated using Equation 1, which measures the cosine distance between the facial expression embedding $\vec{e}$ and the positive reference embedding $\vec{r}_{\text{app}}$, as well as the distance between $\vec{e}$ and the negative reference embedding $\vec{r}_{\text{inapp}}$.

1. Embedding facial expressions and contextual information into a shared vector space.
2. Using contrasting reference points to differentiate between appropriate and inappropriate behaviors in different social contexts (see figure 2).

This approach allows for flexible detection of social appropriateness across different contexts by adjusting reference sentences. We aim to demonstrate both the effectiveness of this method and its sensitivity to environmental context changes.

To evaluate our approach, we conduct two main experiments: In our first experiment, we investigate how changing reference points affects facial description embeddings using six basic emotions. We explore this by testing two types of contexts: neutral and negative. This experiment allows us to examine how the embedding model's understanding of appropriate and inappropriate interactions shifts when the context changes from neutral to negative. By doing so, we aim to demonstrate the flexibility of our approach in capturing nuanced social cues across varying situations.

Our second experiment applies our method to a real-world dataset, focusing on the impact of environmental context changes on social appropriateness classification. We specifically examine how altering the environmental context from the original data collection setting affects the model's performance. This experiment

is designed to highlight the context-dependent nature of social appropriateness and emphasize the importance of aligning the reference context with the actual interaction environment.

These experiments are designed to establish:

1. The capability of embedding models to capture nuances of social appropriateness across different contexts.
2. The sensitivity of our approach to changes in environmental context and its effect on classification performance.

The following subsections provide detailed descriptions of our embedding-based method and the experimental setup used to validate it.

4.1 Social Congruency in Embedding Models

The facial expression frames, consisting of pairs of face images at times t-1 (onset) and t (apex), were submitted to the GPT-4(V) language model via its API. The model was prompted to describe the current emotional state of the person, identifying subtle changes in facial expressions between the onset and apex frames. Example outputs showcasing the model's ability to detect expression shifts are presented in Table 1.

Table 1: Example outputs from GPT-4(V) illustrating detected changes in facial expressions

Facial Expression Descriptions by GPT4(V)
The person seems to be exhibiting a slightly more neutral expression.
The person displays a neutral expression with a slight increase in concern or confusion.
The person appears to be concerned or worried, with slightly furrowed brows and a subtle frown.
The person appears more concerned or worried, with a slight tightening of the eyebrows.
The person seems slightly more concerned or worried, displaying increased tension around the eyes and mouth.
The person wears a neutral expression with a slight increase in alertness or concern.

To establish a relative scale in the embedding space, two contrasting reference sentences were introduced. For the MMI Facial Expression Database, the reference sentences represented appropriate and inappropriate interactions in a neutral setting. Additionally, a negative context ("you are giving bad news to a person") was used to explore how facial expression embeddings can have different score measures when the context changes from neutral to negative (see Table 2).

Table 2: Contrasting reference sentences for MMI and our datasets. Green text indicates correct context; others are for testing wrong contexts.

Dataset	Appropriate Reference	Inappropriate Reference
MMI Database	"This is an appropriate interaction" / "You are giving bad news to a person, this is a socially appropriate interaction"	"This is an inappropriate interaction" / "You are giving bad news to a person, this is a socially inappropriate interaction"
In-the-Wild	"You are a receptionist in the library, giving instructions to people, and this this is a socially appropriate interaction" / "You are at the entrance of a museum, greeting people, and you are having a socially appropriate interaction" / "You are giving bad news news to a person, this is a socially appropriate interaction" / "You are at a at a funeral, interacting with people, and you are having a socially appropriate interaction"	"You are a receptionist in the library, giving instructions to people, and this is a socially inappropriate interaction" / "You are at the entrance of a museum, greeting people, and you are having a socially inappropriate interaction" / "You are giving bad news to a person, this is a socially inappropriate interaction" / "You are at a funeral, interacting with people, and you are having a socially inappropriate interaction"

To evaluate the appropriateness of facial expressions, embeddings for each description generated by GPT-4(V) were calculated using the 'ada-2' model. The cosine distance between these embeddings and those of predefined reference sentences was then determined [23]. The appropriateness score was computed as follows:

$$\text{Score} = d(\vec{e}, r_{\text{inapp}}) - d(\vec{e}, \vec{r}_{\text{app}}) \tag{1}$$

In this equation, $\vec{e}$ represents the embedding of the facial expression description, $\vec{r}_{\text{app}}$ is the embedding of the positive reference point, and $\vec{r}_{\text{inapp}}$ is the embedding of the negative reference point. The function $d(\cdot, \cdot)$ computes the cosine distance between two embeddings (see figure 2).

A higher score indicates that facial expression aligns more with the positive reference point, while a lower score suggests a closer alignment with the negative reference point.

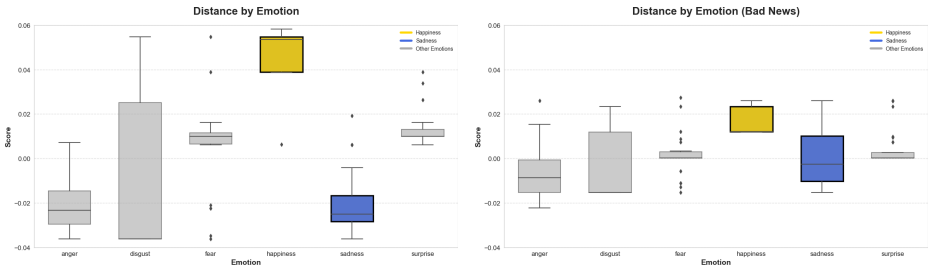
4.2 Analyzing Social Appropriateness in Real-World Contexts

Figure 2 illustrates the process for calculating the social appropriateness score of a facial expression within a given context. For classification purposes, the scores are converted into binary values: positive scores are mapped to 1 (socially appropriate) and negative scores to 0 (socially inappropriate).

To evaluate the model's performance, we primarily use F1 score. The F1 score is preferred because it provides a balance between precision and recall.

We tested the model across different reference points/contexts, as detailed in Table 2, and also presented the confusion matrix for each context. The selected contexts are: a library (the actual context in which the data was collected), a museum, receiving bad news, and a funeral (both representing negative emotional contexts). These diverse contexts were chosen to encompass a range of social norms and appropriateness criteria, showing the effects of these variations on the classification score.

5 Results



(a) Scores for emotions when using the neutral context that "This is a appropriate/inappropriate interaction"

(b) Scores for emotions when using the negative context that "You are giving bad news to a person"

Fig. 3: Scores of different facial description (emotions) sentences with two different contexts embeddings highlighting the two contrasting emotions happiness and sadness

In this section, we present the findings from our analysis using the MMI and in-the-wild datasets under different contextual settings.

For the MMI dataset, Figure 3 illustrates the scores for different emotional expressions when embedded in two distinct contexts. Specifically, Figure 3b shows the scores for the negative context, where the context involves giving bad news, while Figure 3a represents the neutral context, where the interaction is simply appropriate or inappropriate with no negative or positive affect associated with it.

The y-axis in these figures represents the appropriateness score of emotional expressions within the given context, with higher scores indicating greater appropriateness.

In Figure 3a, depicting the neutral context of "This is an appropriate/inappropriate interaction," we observe a contrast between happiness and sadness. The happiness box plot is positioned significantly higher on the scale (mean score: 0.047), with its entire interquartile range above zero. Conversely, the sadness box plot is situated lower (mean score: -0.021), with most of its range below zero.

Figure 3b presents a different pattern for the negative context of "You are giving bad news to a person." Here, the disparity between happiness and sadness is reduced. The happiness box plot has shifted downward (mean score: 0.017) compared to the neutral context, while the sadness box plot has moved upward (mean score: 0.0004). This shift results in an overlap between the two emotions' ranges, with their median lines closer together.

To further evaluate the performance of our proposed approach, we assessed the scores against labeled data from an in-the-wild dataset.

Table 3 presents the confusion matrices and F1 scores for different contextual settings. The library context achieved the highest F1 score of 79%. In this context, 569 instances were correctly identified as appropriate (true positives), and 182 were correctly identified as inappropriate (true negatives). The museum context followed with an F1 score of 64%. In the museum setting, true negatives increased from 182 to 191 compared to the library context, while true positives decreased to 458.

Table 3: Confusion Matrices and F1 Scores for Different Contexts

	Library			**Museum**			**Funeral**			**Bad News**		
	P	**N**	**F1**	**P**	**N**	**F1**	**P**	**N**	**F1**	**P**	**N**	**F1**
Actual P	569	152	**79%**	458	263	64%	438	283	56%	395	326	53%
Actual N	57	182		48	191		172	67		152	87	

The contexts of attending a funeral and delivering bad news yielded F1 scores of 56% and 53%, respectively. In the funeral context, there were 172 false negatives (inappropriate behaviors misclassified as appropriate), higher than in the library and museum contexts. The bad news context showed the lowest number of true positives (395) and a high number of false positives (326).

6 Discussion

The results obtained from both the MMI datset and the in-the-wild dataset provide valuable insights into the role of context in interpreting and perceiving facial expressions. The findings underscore the importance of considering contextual factors in the development of affect recognition systems and highlight the potential for improving human-robot interaction through context-aware affect recognition.

One of the key observations from the MMI Database is the impact of context on the interpretation of facial expressions. The results show how the same facial expression can be interpreted differently through different context integrations. For instance, as can be seen in Figure 3b, expressions typically associated with positive emotions, such as happiness (yellow bar), were perceived by the embeddings model as more negative (lower score) when framed in the context of delivering

bad news. Conversely, the same expressions (yellow bar) were interpreted as more positive (higher score) in a more general context as can be seen in Figure 3a. This finding demonstrates that the perceived positivity of an expression is not solely determined by the expression itself but is heavily influenced by the situational context in which it is expressed.

Our findings from the in-the-wild dataset underscore the critical role of context in emotion recognition. We observed a notable decline in classification accuracy across different contexts. This performance degradation, observed as our model was applied to contexts increasingly dissimilar from the original library setting, accurately captures the nuanced nature of social interactions. It reflects Goffman's concept of "frame analysis" [7], which posits that social interactions are interpreted within specific contextual frameworks. In essence, behavior deemed appropriate in one setting may be inappropriate in another.

The observed drop in F1 scores across contexts highlights the challenges in developing universally applicable models for social behavior classification. This limitation echoes findings in cross-cultural psychology studies, which demonstrate the variability of social norms across different environments and cultures [14]. However, it also underscores a key strength of our proposed method: the need for affective computing systems to adapt easily to the context in which they are deployed.

Our experiments with embedding models, which are trained on extensive datasets, reveal their ability to make connections between appropriate behaviors and physical spaces. These models effectively distinguish between formal settings (e.g., library) and more relaxed environments (e.g., museum). This distinction is further exemplified by our funeral scenario, where social norms are closely tied to the physical location. Our results demonstrate that embedding models can effectively capture these context-specific norms [31,10,24,29].

Notably, even with the use of general contexts like "library" or "museum," we observed significant changes in performance. When slightly altering the context from the actual dataset setting (library to museum), the F1 scores for classifying appropriateness decreased from 79% to 64%.

Despite our promising findings, the deployment of these technologies in the real world poses significant challenges. Firstly, the variability in human responses necessitates a highly adaptive model capable of interpreting a wide array of emotional and verbal cues. Additionally, ethical considerations surrounding privacy and consent, particularly in public settings, require meticulous consideration and adherence to regulations.

7 Conclusion

This work highlights the critical role of contextual information in accurately interpreting social situations, particularly in the domain of HRI. By leveraging sentence embeddings and reference points derived from social scenarios, our proposed approach enhances the ability of AI systems to capture the nuances and contextual cues associated with emotional expressions and social interactions.

The results demonstrate that our method effectively distinguishes between socially appropriate and inappropriate interactions when applied to the same context as the data collection setting, achieving an F1 score of 79% in the library context. However, as the context shifts, the performance deteriorates, with F1 scores dropping to 64% in the museum context, 53% in the delivering bad news context, and 56% in the funeral context.

Our work contributes to improving context understanding in social HRI which is essential for creating adaptive, and socially aware robots. Improvements in context awareness has the potential to enhance user experience, and the overall naturalness of the interaction and eventually contribute to the successful integration of robots into various social settings.

References

1. Amir, S., Wallace, B. C., Lyu, H., and Silva, P. C. M. J. Modelling context with user embeddings for sarcasm detection in social media. *arXiv preprint arXiv:1607.00976* (2016).
2. Anwar, S., Beg, M. O., Saleem, K., Ahmed, Z., Javed, A. R., and Tariq, U. Social relationship analysis using state-of-the-art embeddings. *ACM Transactions on Asian and Low-Resource Language Information Processing 22*, 5 (2023), 1–21.
3. Brown, P., and Levinson, S. C. *Politeness: Some Universals in Language Usage*. Cambridge University Press, 2000.
4. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).
5. Charlesworth, T. E., and Banaji, M. R. Word embeddings reveal social group attitudes and stereotypes in large language corpora, 2022.
6. Dautenhahn, K. Socially intelligent robots: dimensions of human–robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences 362*, 1480 (2007), 679–704.
7. Denzin, N. K., and Keller, C. M. Frame analysis reconsidered, 1981.
8. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
9. Ekman, P. An argument for basic emotions. *Cognition & Emotion 6*, 3-4 (1992), 169–200.
10. Gandhi, K., Fränken, J.-P., Gerstenberg, T., and Goodman, N. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems 36* (2024).
11. Hovy, D., and Fornaciari, T. Increasing in-class similarity by retrofitting embeddings with demographic information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (2018), pp. 671–677.
12. Lemaignan, S., Andriella, A., Ferrini, L., Juricic, L., Mohamed, Y., and Ros, R. Social embeddings: Concept and initial investigation. *Open Research Europe 4*, 63 (2024), 63.
13. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

14. Matsumoto, D. *The handbook of culture and psychology.* Oxford University Press, 2001.
15. Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (2013), vol. 26.
17. Mittal, T., Bera, A., and Manocha, D. Multimodal and context-aware emotion perception model with multiplicative fusion. *IEEE MultiMedia 28*, 2 (2021), 67–75.
18. Mohamed, Y., Ballardini, G., Parreira, M. T., Lemaignan, S., and Leite, I. Automatic frustration detection using thermal imaging. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2022), IEEE, pp. 451–459.
19. Mohamed, Y., Güneysu, A., Lemaignan, S., and Leite, I. Multi-modal affect detection using thermal and optical imaging in a gamified robotic exercise. *International Journal of Social Robotics* (2023), 1–17.
20. Pantic, M., Rothkrantz, L. J., and Koppelaar, H. Context-sensitive gesture recognition. In *Proceedings of the International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction* (1998), Springer, pp. 151–162.
21. Pantic, M., Valstar, M., Rademaker, R., and Maat, L. Web-based database for facial expression analysis. In *2005 IEEE International Conference on Multimedia and Expo* (2005), pp. 5 pp.–.
22. Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2014), pp. 1532–1543.
23. Reimers, N., and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.
24. Sap, M., Rashkin, H., Chen, D., LeBras, R., and Choi, Y. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728* (2019).
25. Stiber, M., Taylor, R., and Huang, C.-M. Modeling human response to robot errors for timely error detection. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2022), IEEE, pp. 676–683.
26. Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., and Schröder, M. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing 3*, 1 (2012), 69–87.
27. Wang, Y., Li, M., Liu, S., and Li, M. Context-aware emotion recognition networks. *arXiv preprint arXiv:1908.05913* (2019).
28. Ye, J., and Skiena, S. The secret lives of names? name embeddings from social media. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2019), pp. 3000–3008.
29. Yuan, Y., Tang, K., Shen, J., Zhang, M., and Wang, C. Measuring social norms of large language models. *arXiv preprint arXiv:2404.02491* (2024).
30. Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*, 1 (2009), 39–58.
31. Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D. Can large language models transform computational social science? *Computational Linguistics 50*, 1 (2024), 237–291.