RESEARCH ARTICLE

Social Embeddings: Concept and Initial Investigation

[version 1; peer review: 2 approved with reservations]

Séverin Lemaignan[1], Antonio Andriella[1], Lorenzo Ferrini[1], Luka Juricic[1], Youssef Mohamed[2], Raquel Ros[1]

[1]PAL Robotics, Barcelona, Spain
[2]KTH Royal Institute of Technology, Stockholm, Stockholm County, Sweden

v1 **First published:** 11 Apr 2024, **4**:63
https://doi.org/10.12688/openreseurope.17296.1
Latest published: 11 Apr 2024, **4**:63
https://doi.org/10.12688/openreseurope.17296.1

Abstract

We introduce *social embeddings* as a compact, yet semantics-preserving, mathematical representation of social situations. Social embeddings are constructed by leveraging pre-trained large language models: we automatically generate a textual description of the social environment of a robot, and use pre-trained text embeddings to generate a vector representation of the social scene. The article presents the details of the methodology, and analyses key properties of these embeddings, including their ability to measure social 'similarity'. We argue that social embeddings are a quantitative pseudo-metric for social situations, we demonstrate their operationalization on actual social robots, and discuss their potential applications.

Keywords

Social representation, embeddings, machine learning, human-robot interaction, AI, digital humanities

Open Peer Review

Approval Status ? ?

	1	2
version 1 11 Apr 2024	? view	? view

1. **Bing Li**, CNRS University of Lille, France, France
   Tatjana A. Nazir, University Lille, Lille, France
2. **Carlo Mazzola**, Istituto Italiano di Tecnologia, Genoa, Italy

Any reports and responses or comments on the article can be found at the end of the article.

H2020 This article is included in the Horizon 2020 gateway.

This article is included in the Marie-Sklodowska-Curie Actions (MSCA) gateway.

Corresponding author: Séverin Lemaignan (severin.lemaignan@pal-robotics.com)

Author roles: Lemaignan S: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Software, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Andriella A**: Conceptualization; **Ferrini L**: Conceptualization, Investigation; **Juricic L**: Conceptualization; **Mohamed Y**: Investigation, Software; **Ros R**: Conceptualization

Competing interests: No competing interests were disclosed.

Grant information: This work has been funded by the EU H2020 ACCIO TecnioSpring INDUSTRY (grant agreement no. 801342, project TALBOT). Additional funding has been recevied from the EU H2020 PERSEO project (grant agreement no. 955778).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Lemaignan S, Andriella A, Ferrini L *et al.* **Social Embeddings: Concept and Initial Investigation [version 1; peer review: 2 approved with reservations]** Open Research Europe 2024, **4**:63 https://doi.org/10.12688/openreseurope.17296.1

First published: 11 Apr 2024, **4**:63 https://doi.org/10.12688/openreseurope.17296.1

1 Introduction

1.1 Social situations and social robotics

The basic idea of *social robots* refers to robots that are situated in a human social environment. In this context, we expect social robots to exhibit *social awareness*, i.e. to appraise and maintain a model of the social situation in which they are embedded. Depending on the role of the robot, this might include understanding who is present, who is interacting with whom, which are the resulting groups, what are the in-group roles, etc.

Social awareness, as a socio-cognitive skill, is essential for the robot to e.g. act in a context-sensitive manner; reason and apply social norms (for instance, do not navigate in the middle of a group, or do not suddenly interrupt a conversation); or create proactive social agents (in order to acknowledge and respond to a human who would like to engage with the robot, the robot must first adequately model and recognise the corresponding social situation).

Social situations have indeed been an important object of study in social psychology for a long time [Argyle *et al.* (1981)], with, for instance, the following definition by Grabett [Garbett (1970)]: *a social situation is a temporally and spatially bounded series of events abstracted by the observer from the on-going flow of social life*. When specifically looking at artificial agents like robots, their level of social cognition depends on their ability to identify and interpret the world surrounding them [Szczepanowski *et al.* (2017)], and in particular, correctly interpret transactions of social signals – specific events in which an agent performs a social action aimed at another agent [Pantic *et al.* (2011)]. This process is known as *situational awareness*, of which Endsley [Endsley (1995)] defines three levels:*perception*, *interpretation*, and *evaluation*.

While the *perception* of social signal has been studied in depth in the HRI community (for instance, [Pantic *et al.* (2011)]), and the *evaluation* of the social situation is usually handled as an aspect of the robot's decision making, the *interpretation* of social situations is a difficult problem. It requires to build and maintain a task-appropriate model of the situations, and represent it in such a way that a machine can reason about it. *Embeddings* are such representations.

1.2 Embeddings

In the context of machine learning, we refer to an *embedding* as a real-valued vector representation of a typically much higher dimensionality input. In other words, a representation that encodes high-dimensionality input (for instance, an image) into a lower-dimensional space. Critically, embeddings are trained to encode the relationships and semantic nuances that might exist in the original input space. For instance, two pictures of the same face transformed with an embedding tuned for facial recognition would yield two vectors that are similar to each other (i.e., close to each other for a given metric, usually the cosine distance). As such, the process of embedding not only condenses high-dimensional information into a more manageable form but also captures latent associations that might otherwise remain obscured [Bengio (2009)].

Unsurprisingly, the training of compact yet semantically-rich embeddings has been a very active research topic over the last two decades, yielding exceptional results in machine learning, where the (otherwise high) dimensionality of real-world percepts might turn common machine learning tasks like classification or prediction intractable.

While research on embeddings initially focused on data that would intuitively lend itself well to mathematical transformations (for instance, reducing the dimensionality of an image, represented as an array of pixel intensities, or processing sound), it has since then been discovered that many constructs – physical or not – can also be *embedded* in a low-dimension numerical space, while preserving many of their semantics. One of the landmark achievements in that regard is the work published in 2013 by Mikolov *et al.* – themselves building on previous work spanning another decade. They showed that embeddings can be computed for *words*, also encoding some of their semantic meaning [Mikolov *et al.* (2013)], with the famous example of *embedding('king') - embedding('man') + embedding('woman') ≈ embedding('queen')*. Effectively, a conceptual equivalence of terms, involving semantics related to gender and social role could be transformed into simple mathematical additions and subtractions.

This outcome ushered in a decade of intense research on text representation, ultimately resulting in the current Large Language Models (LLM) like GPT or Llama2. Importantly for this work, these very large pre-trained networks can also be used to compute text-level embeddings, representing a short text as a numerical vector [Reimers & Gurevych (2019), Muennighoff (2022)]. The resulting embeddings can be then used to measure text-relatedness for instance, as in the BEIR benchmark [Thakur *et al.* (2021)].

1.3 Key insight

Combining the above concepts of social situations and text embeddings, we introduce in this paper the idea of *social embeddings*. A *social embedding* is a compact, real-valued numerical representation (a vector) of a social situation, as experienced by an agent immersed in that social environment. Following the general idea of embeddings, social embeddings are designed to encode the *semantics* of the social situation currently experienced by the agent, facilitating the interpretation of the situation. For instance, it could make it straightforward to compare two social situation by simply measuring how similar the two corresponding embeddings are.

The key insight to construct these embeddings is to exploit the social knowledge already encoded in the latent space of large language models. We do so by automatically generating a textual description of the social environment of the robot (using regular perception routines), and by transforming this description into a text embedding via a large language model. By doing so, we effectively construct an *embedding*, i.e., a projection, of the social space into a machine-friendly numerical space.

This article is a first investigation of this idea. In the following sections, we start the exploration of the design space of social embeddings by presenting a simple algorithm to generate scene descriptions and derive embeddings; we analyse and discuss several key characteristics and parameters of the derived embeddings – like their application as a quantitative *social distance*; and we discuss several promising directions for follow-up research.

1.4 Motivating example

Figure 1 shows three schematic social situations, described from the egocentric perspective of the yellow character. In Figure 1a, the character is engaged with one person, and another group of two people are chatting; in Figure 1b, no one is interacting directly with the character (but someone is walking towards her/him), and three people seem to be chatting together on the side; similarly in Figure 1c, someone is walking towards the character, with another group of two people chatting together.

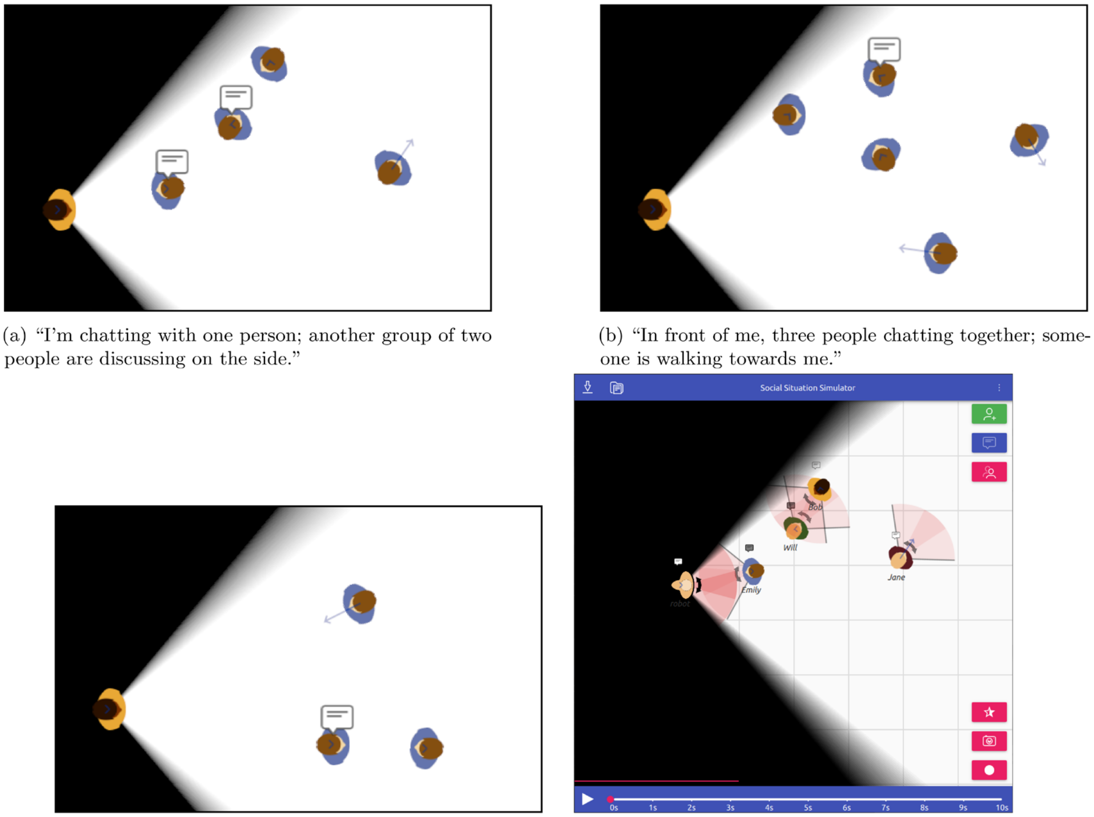
From the perspective of the yellow person the social situations depicted in 1b and 1c are more similar: in both cases, the person is not actively engaged in an interaction yet, but someone seems about to do so. However, a cursory look at the scene could lead to think that 1a and 1b are closer to each other, since they both involve more people than 1c, and most people seem located at the left of the main character.

Accordingly, social embeddings are to be designed so that the situations from Figures 1b and 1c are closer in the embedding space than situations 1a and 1b: by doing so, a robot could correctly compute similarities between social scenes, and e.g. recognise a situation as similar to one it would have previously experienced.

2. Embedding construction

Our approach to embedding construction consists in generating a textual description of the social environment of a situated agent (e.g. a social robot) and computing a text embedding of that description, relying on a pre-trained large language model.

This process entails the following steps: (1) building a model of the social environment of the agent; (2) synthetising a textual description of that environment; (3) computing the text embedding.



(a) "I'm chatting with one person; another group of two people are discussing on the side."

(b) "In front of me, three people chatting together; someone is walking towards me."

(c) "A group of two people are chatting; someone is walking towards me."

(d) Social situation editor interface (representing the same scene as Fig. 1a).

Figure 1. Three simple social situations, described from the perspective of the yellow agent (arrows indicate direction and magnitude of motion). Figure 1d shows the open-source social situation editor used to create the test scenarios.

2.1 Social signals and situation perception

Step 1 relies on the availability of social perception capabilities, both hardware (for sensing) and software (e.g. recognition and tracking of people). For this initial exploration of social embeddings, we focus here on perception capabilities commonly found on social robots. Namely, we build social embedding from the following social signals:

- relative position of agents;
- tracking of agents (the same person is always refered with the same identifier across frames);
- velocity and direction of movement;
- gaze direction (field of view and field of attention);
- (optionally) whether or not agents are currently talking.

These signals are readily available using perception frameworks like the open-source ROS4HRI [Mohamed & Lemaignan (2021)].

Importantly, we adopt a *situated* and *ego-centric* perspective: the scene is described *as seen by an agent, itself embedded in the social situation*. This choice is driven by practical considerations: we are designing social embeddings to be a useful tool for a social robot to analyse and recognise the current situation it is embedded in; as such, the robot is itself part of the social situation, and must describe the world from its own perspective. The three scenarios pictured in Figure 1 are 2D top-down representations of such situated and ego-centric perspectives on three social situations.

To evaluate our approach, in this work we also use a custom-made social situation simulator (Figure 1d) to quickly generate synthetic social data as described in Section 3.1.

2.2 Scene description

Step 2 consists in converting the perceived social situation into a textual description. Many approaches are possible. In this work we choose a relatively direct translation of the social signals into a concatenated list of *descriptors* separated by semi-colons. Examples of descriptors are: “X is looking at Y”, if Y lies within the field of attention of X; or “X is walking towards me”, if the dot product of the motion and relative position vectors of X is close to −1.

Specifically, we compute the following descriptors:

- *X is close to Y* (< 1.5*m*)
- *X is not far from Y* (< 4*m* and facing each other)
- *X is walking towards me*
- *X is walking away from me*
- *X is passing by*
- *X is looking at Y*
- *X and Y are looking at each other*
- *X is talking* (if speech detection is available)

Algorithm 1 details the computation of these descriptors. The name placeholders in the resulting descriptors are replaced by random names (yet keeping the same name for a given

Algorithm 1: Generation of social situation descriptions

```
input : $(X_i, \dot{X}_i, \Theta_i, isTalking_i)$ with $i \in Agents$
input : $ego$, the index of the agent whose perspective we adopt
output : a set of textual descriptors
CLOSE ← 1.5$m$; MEDIUM ← 3$m$; FAR ← 4$m$
$\delta_{i,j} \leftarrow ||X_i - X_j||$
FoV ← 100°; FoA ← 30°
descriptors = ∅
/* j sees i? */
func isVisibleBy i, j
    return $|arctan(X_i - X_j) - \Theta_j| < FoV/2$
/* i looks at j and j is the closest to i? */
func isLookingAt i, j
    $\Lambda = \{k, |arctan(X_k - X_i) - \Theta_i| < FoA/2\}$
    return $j == argmin_k(\delta_{i,k}, k \in \Lambda)$
func isFacing i, j
    return isVisibleBy(i, j) ∧ isVisibleBy(j, i)
func relativeDistance i, j
    if $\delta_{i,j} < CLOSE$ then return “{i} is close to {j}”
    if $(CLOSE < \delta_{i,j} < FAR) \wedge isFacing(i, j)$ then
        return “{i} is not far from {j}”
func relativeMotion i, ego
    if $CLOSE < \delta_{i,ego} < FAR$ then
        $angle \leftarrow \frac{\dot{X}_i - \dot{X}_{ego}}{\|\dot{X}_i - \dot{X}_{ego}\|} \cdot \frac{X_i - X_{ego}}{\|X_i - X_{ego}\|}$
        switch angle do
            case $angle < -0.7$ do
                return “{i} is walking towards me”
            case $angle > 0.7$ do
                return “{i} is walking away from me”
            case $|angle| < 0.3$ do
                return “{i} is passing by”
for Agent $i \neq ego$ do
    if isVisibleBy(i, ego) then
        descriptors ← relativeDistance(i, ego) descriptors ← relativeMotion(i, ego) if $isTalking_i$ then
            descriptors ← “{i}istalking”
            for Agent $j \neq i$ do
                if isVisibleBy(j, i) ∧ isVisibleBy(j, ego) then
                    if $isLookingAt(i, j) \wedge \delta_{i,j} < MEDIUM$ then
                        if isLookingAt(j, i) then
                            descriptors ← “{i}and{j}arelookingateach other”else
                                descriptors ← “{j}islookingat{i}”
                return descriptors
```

agent). By applying this algorithm, the situation pictured in Figure 1a leads to the following description:

"Emily is talking; Emily is close to me; Emily and I are looking at each other; Will is talking; Bob and Will are looking at each other; Bob is not far from me; Bob is looking at me"

Likewise, Figure 1b is described as:

"Bob is talking; Bob is looking at Emily; Bob is not far from me; Bob is looking at me; Will is walking towards me; Will and I are looking at each other; Will is not far from me"

Finally, Figure 1c is described as:

"Violet is walking towards me; Violet is not far from me; Violet is looking at me; Emily is talking; Emily and Will are looking at each other; Will and I are looking at each other; Will is not far from me".

As the aim of the embedding is to compactly represent the semantic of the social situation, the pragmatics of the description, and in particular, the order of the descriptors in the final description, should have little influence on the resulting vector. Likewise, the randomly drawn names should not impact the semantic proximity of two similar social situations within this context. We evaluate these properties in Section 3.

2.3 Embedding generation

The last step consists in computing a text embedding of the description. We first add some task context by wrapping the description in the following text: *"This is the description of a social setting with a few people: [generated descriptions]. We want to generate a good description of the situation."*, before transforming the full text with a text embedder.

We use `langchain`[1] to abstract the call to the text embedder. In the following sections we primarily use Meta's Llama2 13B text embedder, while comparing it to OpenAI's `ada-002` in Section 3.5.1. Moreover, in Section 3.5.2, we discuss in greater details the use and impact of the added task context. Our results show that, depending on the text embedding model, task context do or do not have significant impact on the representation of social situations.

3 Evaluation on synthetic data

To be useful, social embeddings must exhibit at least the following key properties:

- **pragmatics invariance**: if we want the represent the *semantics* of a social scene, the pragmatics of the description (exact wording, order in which we describe the scene, etc.) should only have a minimal influence;
- **semantic similarity**: two social situations that are similar should result in embeddings that are close to each others in the embedding space;
- **continuity**: small changes in the social situation should result in correspondingly small variations in the embedding space.

As our methodology does not depend whether the data is synthetic or acquired in real-world conditions, we perform first an in-depth evaluation on simulated social data, and we demonstrate how the same methodology can be applied to real-world data in Section 4.

3.1 Generation of synthetic social data

We have developed an open-source social interactions 'editor'[2], (Figure 1d) that we use to design small interaction situations. The tool allows the manual creation of short animations of 2D characters moving around, forming groups, and talking. Scenarios are generated by creating keyframes at various points in the timeline, with the tool interpolating the position, orientation and velocity of the characters between frames. Characters can also be marked pairwise as 'interacting' with each other (this particular information is not used to generate embeddings descriptions; it could however be used in future work to e.g. train a engagement detector).

The editor saves scenarios as JSON files; it can also generate snapshots and videoclips of the interactions, as viewed from any character's perspective (see examples in Figure 1). These video-clips can then be used for e.g. online studies.

The output JSON files are then post-processed by a set of helper scripts (also open-source) to perform the following steps:

1. generate, for each agent in the scene, and at every time point, *situation descriptions* (using the algorithm presented in Section 2.2).

   The descriptions are normalised such that a similar social situation, viewed by a different agent, will have the exact same description (e.g., same descriptor ordering); the names of agents are represented as *slots* like `{A}`, `{B}`,... so that new identical descriptions with different names can later be generated.

2. compute the embeddings, optionally randomising the order of the descriptions, and/or randomising names to create variations of the same situations.

The end result is a CSV file containing a list of descriptions, a group ID associating together the variations of the same description (if they are generated), and the embedding vector

[1] https://github.com/langchain-ai/langchain

[2] source attached to submission; final manuscript will include link to code repository

(the dimension of the embedding vector depends on the text embedding model; for instance, $d = 5120$ for Meta's Llama2 model).

In the following sections we use two short scenarios created with this tool to extract a set of social situations. The two scenarios[3] last a total of 48 seconds and involve five people in a variety of changing group situations: people move from one group to the other, sometimes interact, sometimes not. Figures 1a to 1c are snapshots from this dataset.

These scenarios have been created *ex-nihilo* to generate a broad range of different social situations – they do not attempt to faithfully represent actual human social interactions; their purpose is instead to represent prototypical interactions that should test the representational capabilities of the social embeddings presented in this work.

The scenarios are sampled at 2Hz, and each scene is described from the perspective of each of the five persons, leading to 48 × 2 × 5 = 480 unique situations. Once duplicate descriptions are removed, we obtain 123 unique scene descriptions.

Note that the data analysis scripts used below are attached to the submission as a Python notebook.

3.2 Invariance to pragmatics

We first investigate whether our embeddings actually encode the semantics of the social situation, and not merely the lexical structure of the description.

To this end, and for each situation description, we create one random variant by randomising the order of the descriptors. We then compute the cosine distance between the original description embedding and its variant.

We compare the resulting distribution of distances to the distances of the original description to every other descriptions in the scenario. Figure 2 shows the resulting density distributions, with Table 1 reporting additional descriptive statistics.

For comparison, we perform a similar computation directly on the textual descriptions (hereafter the *textual space*), using the Levenshtein distance as a string similarity measure.

We see on Figure 2, left, that random variations of descriptions end up very close to their original, canonical descriptions (distances close to zero). In contrast, using string similarity as metric (Figure 2, right), the distances between pairs (canonical description, variants) are spread over, where two descriptions of the *same* scene are not typically close to each other.

This first result indicate that our embeddings do encode at least some semantics of the scene description, independently of the exact pragmatics of the description.

3.3 Social similarity

The main aim of social embeddings is to provide a reliable metric of social situation similarity. While we intuitively understand it, the 'similarity' of two social situations is actually not easily quantified, and to the best of our knowledge, no tool exists to automatically measure how similar two social situations are, beyond manual annotations by experts.

[3] included in manuscript attachment

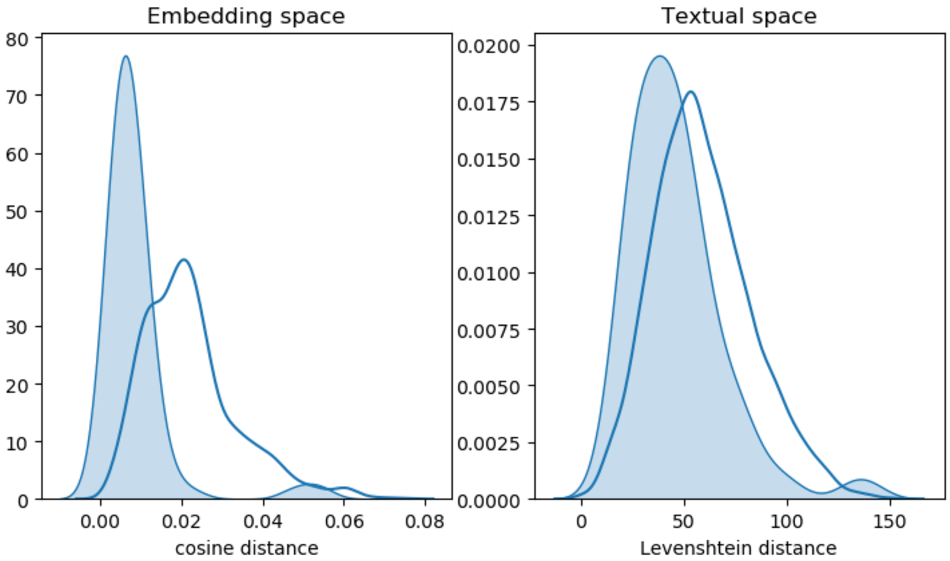


Figure 2. Density distribution of the distances in embedding space (cosine distance, left) and textual space (Levenshtein distance, right). The shaded area correspond to measured distances between each social scene descriptions and one of their variants.

Table 1. Distances between representations, embedding space vs textual space. The cosine distance is used to compute the distance between embeddings; the Levenshtein distance is used to compute similarity between textual descriptions.

	n	**Embedding space**		**Textual space**	
		mean	***stddev***	***mean***	***stddev***
→ to variant	90	0.0084	0.0089	46.1	22.66
→ to other descriptions	4005	0.022	0.012	59.21	23.63

Because it would be difficult for people to rate on an absolute scale whether two situations are similar or not, we approached the problem through pairwise comparisons (akin to A/B testing): 'are two randomly-chosen situations more similar to each other, than two other randomly-chosen situations?'

To assess whether our embeddings correctly encode the similarity of social scene, we therefore designed an online study where participants were shown a short 2-seconds video clip of a randomly-picked social situation animation (in same visual style as Figure 1). This first situation was referred to as the 'reference situation'. Two additional clips (video A and video B) where displayed below, and the participants were given the following options: select the clip (A or B) which resembled the most to the reference situation; indicate that both clips where very similar to the reference situations; or that both clips were very different from the reference situation (see Figure 3).

We collected 893 such pairwise comparisons. Out of these, 92 were double-coded by at least two independent experts, with 74 achieving agreement (inter-rater agreement measured with Kripendorff's alpha: 0.68). In the following, we only keep these $n = 74$ comparisons.

We then ran the same comparisons using the cosine distance in the embedding space. We implemented a simple threshold-based predictor over the four classes (A closer to reference; B closer to reference; A and B both similar to reference; A and B both different to reference). We used the 20th and 80th quartile of the full distribution of distances as thresholds for 'similar' and 'different', respectively.

Using Llama2, the 4-classes accuracy of the predictor is measured at 60.8% (chance, computed by setting random distances between embeddings, is at 33.3%). Figure 4 shows the full confusion matrix of this heuristic-based classifier. Note that the *B closer than A* row has zero true label, as we swapped videos A and B (and their votes) during the data analysis to always have video A reported as more similar to the reference than video B in order to simplify the interpretation.

These initial results show that, while not perfect, our social embeddings, derived from simple scene descriptions, already perform significantly better than chance when trying to quantitatively assess the similarity of social situations.

While this first study has important limitations (simplified model of social interactions, no precise measurement of how similar or dissimilar situations are, etc.), it does support the general validity of the approach.

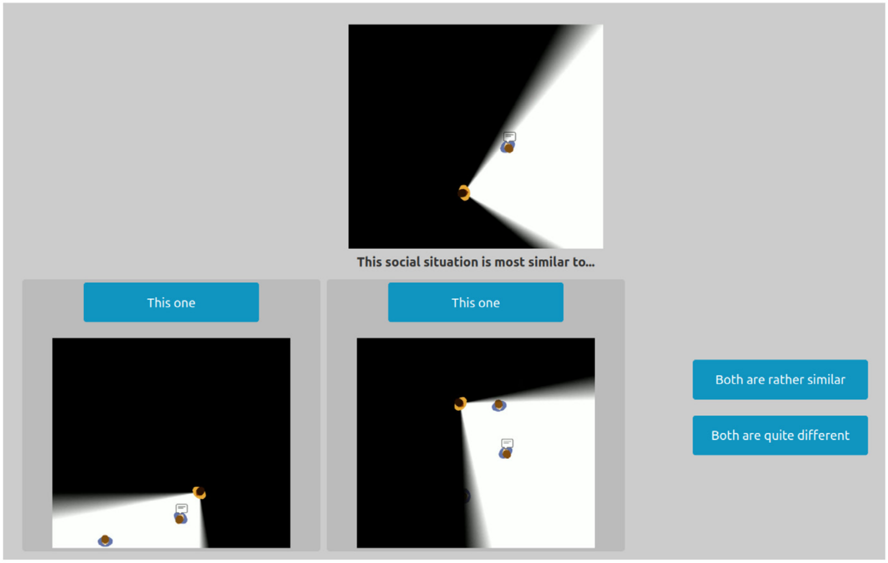
3.4 Continuity

The third desirable property of our embedding is *continuity*: small changes in the social environment should lead to similarly small changes in the embedding space. Formally, if we denotes $\mathcal{S}$ the set of social situations, the embedding process $f : \mathcal{S} \rightarrow [0; 1]^n$ is continuous iff $\lim_{x \to c} f(x) = f(c)$ for any x and c in $\mathcal{S}$.

This formulation is theoretical, as it assumes a metric of social similarity between x and c that is not available (at least not in a quantitative way). Regardless, our embeddings are by construction *not* continuous, as the textual descriptions we generate are built from a discrete set of descriptors, and the addition or removal of one descriptor will cause in practice a discontinuity in the embedding space.

Future work might investigate into how this could be mitigated, by e.g. redefining the concept of continuity for social situations. Nevertheless, looking at the sequence of social embeddings during a continuous interaction provides interesting insights. For instance, Figure 5 represents the distance between consecutive embeddings of the social situation experienced by one of the agents in one of our synthetic test scenario. In effect, this graph represents the rate of social change experienced by this agent over the 39 seconds of the scenario.

For example, in this instance, the agent was in a stable social situation between seconds 20 and 28, while it experienced rapid and significant change of social environment around second 10.



This social situation is most similar to...

Figure 3. Screenshot of the online study setup.

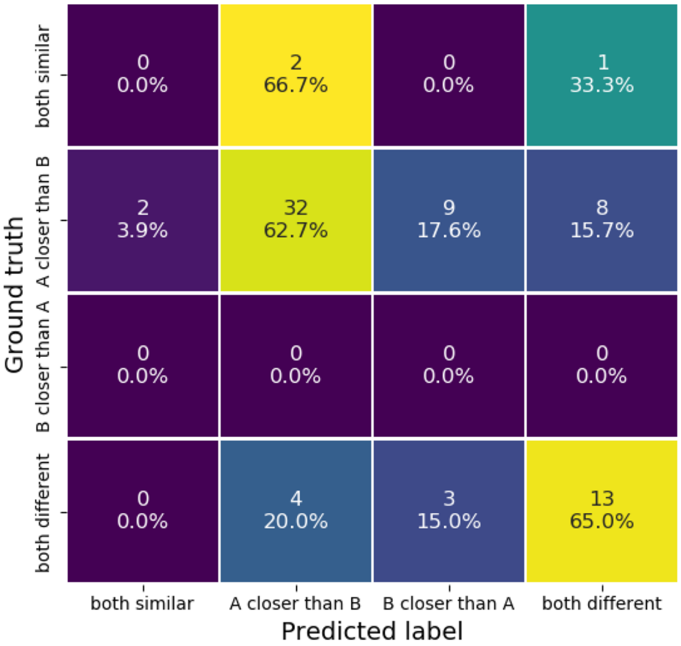


Figure 4. Confusion matrix between the human baseline of similarity of social situations vs similarity predicted using distances in embedding space. Both absolute number of situations, and number of situation normalized with respect to the true labels, i.e. *y*-axis, are reported; see text for ‘B closer than A’ zero true labels.

3.5 Other hyper-parameters

The design space of social embeddings, as we define them, is large, and many parameters might influence the final quality of the embed representation. In this section, we examine some of those hyper-parameters, namely: the model used for text embedding; the role of task context; the generation of ego-centric versus allo-centric descriptions; and the use of human names versus abstract identifiers.

3.5.1 Comparison between Large Language Models. Different large language models have different text embedding space, which might be more or less favourable to encoding the

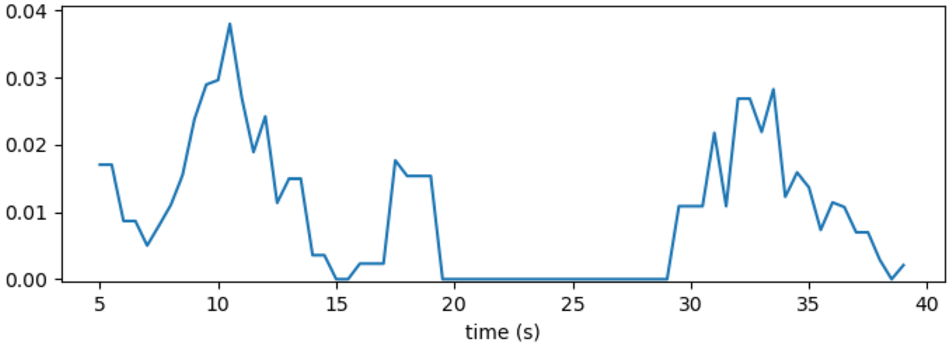


Figure 5. Distances between two consecutive social situations viewed by one agent in the test scenario 2. Embeddings are sampled at 2Hz, and distances smoothed over a 2 seconds rolling window.

semantic of social situations. We compare the Meta Llama2 13B model used so far to OpenAI's `text-embedding-ada-002` model.

At the time of writing, Llama2 is the leading open-source large language model, and while a 70B version of this model is available (and might yield even better results), we settled for the 13B with 4 bits quantization applied, so that we could generate the embeddings ourselves, on consumer-grade, CPU-only hardware (on an Intel i7 11th, it took on average 17 seconds per embedding). Indeed, the ability to generate offline embeddings on hardware found on off-the-self robots is critical to ensure future privacy-concious applications.

OpenAI `ada-002` performs text embedding in a $d = 1536$ dimensional space, versus $d = 5120$ for Llama2. Performing the same computations as in Section 3.3 with the `ada-002` embeddings yields an overall classification accuracy of 56.8%, slightly below the Llama2 results (60.8%). While the results are close to each other, Llama2 seems to perform slightly better on this task, which might be partially explained by the higher dimensionality of the embeddings.

3.5.2 Influence of the description context. As explained in Section 2.3, we computed our social embeddings by inserting the generated scene description in a short task context (*"This is the description of a social setting with a few people: [descriptors]. We want to generate a good description of the situation."*).

The use of such *task context* is suggested by research on LLM prompting (for instance [Kojima *et al.* (2022)]) showing that simple prompt modifications, like adding "Let's think step by step" can very significantly improve the reasoning capabilities of many LLMs.

To better understand the impact of context-setting on the embedding of social situations, we computed social embeddings with four different contexts:

- no context (only the descriptors);
- short context (see text above);
- short context with *"Let's think step by step"* appended at the end;
- longer context that include one example of inference, and example of possible follow-up questions (for instance, after the scene description, the context include the text *"Am I engaged in an interaction right now? if so, with whom?"*).

Running the classification task with the OpenAI model, the accuracy results are respectively 58.1%, 56.8%, 56.8% and 56.8%. It generally appears context setting has little impact on the quality of the representation, and in particular, adding *"Let's think step by step"* does not improve classification results at the level of the text embedding.

In the case of Llama2 13B, the classification results were 31.1%, 60.8% and 56.8% (long context omitted due to prohibitively long computation costs). The major drop of accuracy when embedding descriptions in Llama2 without context is unexpected, and would require further investigation in the internals of the Llama2 model.

3.5.3 Egocentric vs allocentric descriptions. In Section 2.2, we generate a description of the social environment from the *egocentric* perspective of the agent (first-person perspective). For instance, we generate descriptors like *Mary is walking towards* **me**. This choice is driven by the intuition that social experiences that would be part on the training corpora of large language models would predominantly be narrated from a first person point of view.

The alternative is to describe the scenes from a 'god-like' *allocentric* perspective, where we do not adopt a particular agent's point of view. The first person pronoun is replaced by an arbitrary name, and the previous example would for instance become *Mary is walking towards* **John**.

Running the same classification task using allo-centric description with the OpenAI `ada-002` text embedder yields an overall accuracy of 55.4%, with 63.0% for the Llama2 model.

While the difference is small, it appears that egocentric descriptions yield slightly better social representations.

3.5.4 Use of human names vs abstract identifiers. As presented in Section 2.2, agents present in a given scene are assigned arbitrary human names – only ensuring that the same agent receive the same name in subsequent frames of the same scenario.

The intuition for assigning human names rather than random identifiers (like *A*, *B*, *C*...) is that is might help steering the embeddings towards representing social interactions, by implicitly signaling that the agents are humans.

Re-running the classification task with letter identifiers instead of names however actually slightly improves the classification results (62.2% vs 60.8%). This result might however be an artifact of our classification task. Indeed, participants were asked to compare social situations without any specific information about the identities of the characters drawn on the screen: the characters appearing in the reference video clip might or might not be the same as the ones appearing in the two other videos. However, by using human names, the embed descriptions did encode the identities of the characters across videos. On the contrary, we can hypothesise that, when using letters instead of names, we implicitly signal that the agents should be treated as anonymous entities, resulting in embedding representation more similar to what the human annotators would have experienced.

4 Validation on real-world data

Because the Algorithm 1 is designed to take as input social signals that are commonly found on robots, there is no conceptual difficulty in running the same algorithm on actual hardware. As a proof-of-concept, we recorded a short sequence (1min 42sec) of four people interacting in front of a PAL Robotics ARI robot [Cooper *et al.* (2020)]. The robot is running out-of-the-box a ROS4HRI-compatible pipeline [Mohamed & Lemaignan (2021)] that exposes the location, gaze direction and velocities of the people around the robot. The robot does not support speaker extraction and identification, so we could not uniquely identify who was speaking. Instead we relied on the available voice activity detector to simply record that ‘someone’ was speaking.

Scene descriptions were generated at 3Hz (maximum permissible rate to generate embeddings in real-time from the OpenAI API endpoint), using the same algorithm as Algorithm 1 (except for the {*agent*} *is talking* replaced by *Someone is talking*).

Using these computed embeddings, and similar to Figure 5, Figure 6 represents the rate of social change experienced by the robot. As illustration, snapshots of the robot’s camera have been added before and after two peaks of social change.

5 Discussion

5.1 Generation of descriptions and ethical considerations

The process to generate textual descriptions of social situations is critical to the computation of our social embeddings. In this initial investigation, we took a simple approach consisting in concatenating descriptors directly computed from commonly available robot’s sensing.

We have not yet investigated alternative approaches – or whether all the descriptors that we use are in fact required. Future work should investigate this design choice, investigating possible richer descriptions (e.g. individuals’ expressions, group formation recognition, etc.), and running ablation studies to determine if particular descriptors are redundant.

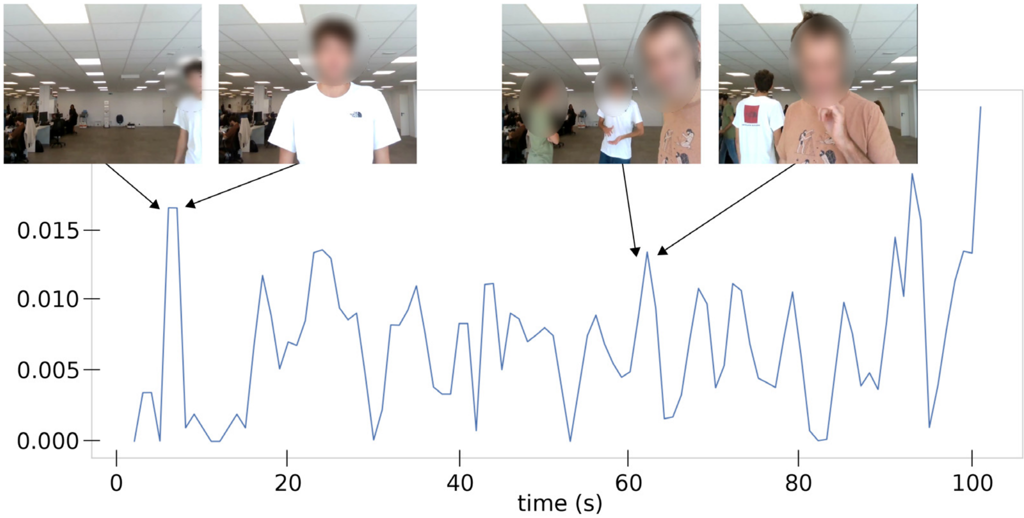


Figure 6. Cosine distances between two consecutive embeddings sampled at 3Hz, and averaged on a 2sec rolling window.

The way descriptions are generated also has ethical implications. Because we rely on the use of LLMs to generate embeddings, we might introduce undesirable biases in the obtained representations. Indeed, as reported by OpenAI[4] for instance, "*the models encode social biases, e.g. via stereotypes or negative sentiment towards certain groups*." Thus, careful attention must be paid when designing the input descriptors to generate the embeddings. In this work we include one potentially biased descriptor: the name identifiers. While they are randomly picked from a gender-balanced list, they are all English names, and do not balance diverse cultural backgrounds. Future work might look into additional sensitive descriptors, like age, gender, task context, that could unwillingly lead to encoding stereotypes. We therefore encourage researchers to evaluate the adequateness and potential impacts of the descriptors introduced in the social embeddings construction based on the context of use.

5.2 Context awareness

The correct interpretation of a social situation is strongly context-based: for example, a person starting to laugh at a robot designed to help visitors in a library likely indicates some inappropriate behaviour, while the same person laughing while the robot is telling a joke would on the contrary indicate a successful interaction.

While we have not explored the representation of context in this first presentation of social embeddings, our approach lends itself well to this kind of extension: assuming one can create a description of the current context of interaction ('We are in the reception hall of an hospital and my role is to welcome visitors'; or 'We are in an office, and people do not want to be disturbed without reason'; etc.), the context textual description could simply be prepended to the scene description, to embed as well contextual information. This represents an important area of future research.

5.3 Latent semantics

In this initial investigation, we have not attempted to uncover the latent semantics encoded in the embeddings. For instance, we can expect that social situations like 'two persons chatting together'; or 'a group of three people walking together'; or 'one single person walking towards the robot'; etc. are all semantically distinct, and, consequently, would belong to distinct regions in the embedding space. Identifying such clusters to broadly characterize the semantic topology of the embedding space would enable not only to measure how similar two social situations are, but also characterize key features of the current situations. This idea is related to e.g. the recent investigation by Sun and Nelson [Sun & Nelson (2023)] on latent semantics of sentence embeddings.

5.4 Fine-tuning

The text encoders used to generate the embeddings were not fine-tuned for the specific task presented in the paper. However, contrastive-learning-based fine-tuning [Hadsell *et al*. (2006)] has shown potential for instruction-based text embeddings [Canal *et al*. (2022)]. Exploring contrastive-learning-based fine-tuning techniques is expected to improve the results reported here.

A major challenge, in addition to the fine-tuning itself, is however the collection of a dataset expressing differences and similarities between social situations. The collection of this type of data is a non-trivial problem. Intuitively, it will require participants labelling social situation as more or less similar, based on their perception and understanding.

6 Conclusion: Social Embeddings as a Social Metric?

We have introduced in this paper the concept of *social embeddings*: compact vector representations of arbitrary social situations, computed by first generating a textual description of the social environment, and then leveraging pre-trained large language models to embed the resulting description into a low-dimensional numerical space. Moreover, we have shown evidence that the resulting embeddings encode some of the social semantics of the original situations, and their topology partially reflects the qualitative similarity observed by humans between given social situations.

By providing a algorithmic bridge between qualitative appraisal and quantitative measurement of similarity, we can naturally question whether social embeddings could in fact be a metric over the 'social space'. Formally, the definition of a metric d over a space implies the following four properties:

1. distance to itself is zero: $d(x, x) = 0$
2. positivity: If $x \neq y$, then $d(x, y) > 0$
3. symmetry: $d(x, y) = d(y, x)$
4. triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$

While properties 1–3 are trivially derived from the computation of the cosine distance between embeddings, the triangle inequality property is difficult to formally prove for text embeddings. An empirical evaluation with 1000 randomly-picked triplets of scenes (x, y, z) shows that in 96.1% of the cases, the inequality $d(x, z) \leq d(x, y) + d(y, z)$ is verified. As such, while they are not formally a metric of social situations, social embeddings appears to behave as a good pseudo-distance for the social space.

While we have made it abundantly clear that much work remains to be done to refine and fully characterize social

[4] https://platform.openai.com/docs/guides/embeddings/limitations-risks

embeddings, equipping robots with such a metric over the social space would open multiple new venues for research and applications. In order to conclude this initial investigation of the potential of social embeddings, we list hereafter some of these mid-term/long-term ideas, hoping to initiate fruitful new lines of research on data-driven social psychology and HRI:

- by clustering a diverse set of social situation embeddings, it should be possible to isolate key prototypical social situations, to which rules, norms, behaviours could be associated. A robot could then recognise when it encounter such situations, and also estimate the *dynamics* of the current situation, i.e. towards what other situation it is headed;
- this ability opens the door to social prediction, letting the robot anticipate future social state (and adjust accordingly its behaviour to avoid undesirable situation, or facilitate desirable ones);
- as vector representations, social encodings lend themselves well to many downstream machine learning tasks, for instance socially-appropriate behaviour generation, or situated interactive machine learning;
- because social embeddings also lend themselves to encoding task and social context, they open new possibilities for context-aware representation and behaviour selection;
- when used to represent changes in the social environment (as in Figure 5 and Figure 6), social embeddings could also become a powerful tool to automatically recognise interaction issues: discontinuities in the embedding space would represent sudden changes in the social environment that might be unexpected in the current context of the robot.

Data availability

Zenodo: Social Embeddings: social simulator and original dataset of sample social situations, https://doi.org/10.5281/zenodo.10623138 (Lemaignan, 2024).

This record contains:

- The source code of a social situation simulator used to record 2D top-down social interactions between people
- Samples of social interactions recorded with the simulator
- An annotated IPython notebook with the various data analysis steps followed in the ’Social Embeddings: Concept and Initial Investigation’ paper
- computed social embeddings for these interactions

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

References

Argyle M, Furnham A, Graham JA: **Social situations.** Cambridge University Press, 1981.
Publisher Full Text

Bengio Y: **Learning deep architectures for AI.** *Found Trend Mach Learn.* 2009; **2**(1): 1–127.
Publisher Full Text

Canal FZ, Müller TR, Matias JC, *et al.*: **A survey on facial emotion recognition techniques: A state-of-the-art literature review.** *Inform Sciences.* 2022; **582**: 593–617.
Publisher Full Text

Cooper S, Di Fava A, Vivas C, *et al.*: **ARI: The social assistive robot and companion.** In: *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN).* IEEE, 2020; 745–751.
Publisher Full Text

Endsley MR: **Toward a Theory of Situation Awareness in Dynamic Systems.** *Hum Factors J.* 1995; **37**(1): 32–64.
Publisher Full Text

Garbett GK: **The Analysis of Social Situations.** *Man.* 1970; **5**(2): 214–227.
Publisher Full Text

Hadsell R, Chopra S, LeCun Y: **Dimensionality reduction by learning an invariant mapping.** In: *2006 Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit (CVPR’ 06).* IEEE, 2006; **2**: 1735–1742.
Publisher Full Text

Kojima T, Gu SS, Reid M, *et al.*: **Large language models are zero-shot reasoners.** *Adv Neural Inf Process Syst.* 2022; **35**: 22199–22213.
Publisher Full Text

Lemaignan S: **Social Embeddings: social simulator and original dataset of sample social situations.** [Data set]. Zenodo. 2024.
http://www.doi.org/10.5281/zenodo.10623138

Mikolov T, Chen K, Corrado G, *et al.*: **Efficient estimation of word representations in vector space.** *arXiv preprint arXiv: 1301.3781.* 2013.
Publisher Full Text

Mohamed Y, Lemaignan S: **ROS for Human-Robot Interaction.** In: *Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems.* 2021.
Publisher Full Text

Muennighoff N: **Sgpt: Gpt sentence embeddings for semantic search.** *arXiv preprint arXiv: 2202.08904.* 2022.
Publisher Full Text

Pantic M, Cowie R, D’Errico F, *et al.*: **Social signal processing: The research agenda.** *Visual Analysis of Humans.* 2011; 511–538.
Publisher Full Text

Reimers N, Gurevych I: **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.** arXiv: 1908.10084 [cs.CL], 2019.
Publisher Full Text

Sun T, Nelson B: **Topological Interpretations of GPT-3.** arXiv: 2308.03565 [cs.CL], 2023.
Publisher Full Text

Szczepanowski R, Gakis MG, Arent K, *et al.*: **Computational Models of Consciousness-Emotion Interactions in Social Robotics: Conceptual Framework.** In: *Cognitive and Computational Neuroscience-Principles, Algorithms and Applications.* IntechOpen, 2017.
Publisher Full Text

Thakur N, Reimers N, Rücklé A, *et al.*: **BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models.** *arXiv preprint arXiv: 2104.08663.* 2021.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 20 December 2024

https://doi.org/10.21956/openreseurope.18694.r47021

?

Carlo Mazzola
Istituto Italiano di Tecnologia, Genoa, Italy

The article “Social Embeddings: Concept and Initial Investigation” introduces the idea of Social Embeddings with a position paper enriched by data analysis and proof of concept. Social Embeddings aims to be a quantitative representation of social situations. They encode the complexity of the social dynamics contained in text descriptions extracted from the visual representation of the scene in a continuous, lower-dimensional space using large-language models (LLM). The paper describes the methodology used to create social embeddings, starting from the design of the scenes with a custom social situation simulator. After the evaluation of synthetic data, the pipeline is deployed and analyzed on the ARI robot as proof of concept of the approach for social robotics.

The paper proposes an interesting application of word embedding analysis to the interpretation of social contexts that robots can perceive multimodally (vision and audio). The topic is currently of high interest in the community at the edge between social robotics, human-robot interaction, and neural networks, and offers an interesting solution for a hard problem such as the quantitative analysis of social dynamics. Overall, it provides a clear description of the methodology and is well-written. Since the authors aim to propose a new methodology, however, I believe the paper should be more strongly rooted in the previous literature on the topic. Moreover, I think the paper needs some improvement to be approved, which I am listing below.

1. The authors state that their article is the first investigation into encoding automatically generated textual descriptions of the social environment in LLM’s generated embeddings to interpret social scenes. However, I see very little analysis of the state-of-the-art (in particular, of very recent literature) for such a statement. Since the paper aims to propose a new methodology, I invite the authors to write a throughout state-of-the-art section on the topics that are connected more closely to their work. For instance, Word Embeddings for Social Analysis, Group Conversation in HRI, Socially Aware Robots...
2. The authors should refer more often to previous methodologies in the literature to strengthen the foundation of their approach. For instance, they should cite relevant works regarding the evaluation of synthetic data and the selection of important scene descriptors. This will provide a clearer context for their methodological choices.

3. Using the same model-based algorithm to generate textual descriptions of the scene may result in repetitive or overly similar descriptions. While the authors' effort to verify whether the embeddings capture purely textual features or deeper semantic information is commendable, the analysis needs further expansion. Moreover, the policy used to determine the degree of similarity between texts or embeddings is not clearly defined and should be made more transparent.
4. Section 3.3, "Social Similarity" should be expanded in the method description to improve reproducibility. The transition from 893 pairwise comparisons to 74 comparisons needs further explanation. Specifically, it is unclear whether the dataset was balanced across the four classes and how this might have affected the analysis. In this respect, authors might want to use different indicators of the classification performance, such as weighted F1 score. Additionally, the threshold policy used to assess similarity should be more thoroughly explained, including the criteria for its selection and its impact on the results. Moreover, the transparency of the authors' approach could (and should) be increased by incorporating visualizations of the embedding space through additional dimensionality reduction techniques, such as PCA. This would allow both readers and developers to better understand the structure of the embeddings and provide clearer insight into how the social dynamics are represented in the reduced space.
5. Validation of real-world data: this section requires further elaboration, with more data and analysis to provide stronger support for the findings. Since Figure 6 does not seem to offer sufficient insight into the effectiveness of the solution for the embodied system, it would be beneficial to conduct an additional analysis, similar to the approach in Section 3.3, using ground truth data to evaluate the model performance. In this regard, have the authors considered using an existing dataset for validation before deploying the model on the robot? This could provide a more robust foundation for the results and offer clearer comparisons.
6. From a theoretical perspective, the distinction between egocentric and allocentric spatial representations does not seem always accurate. In the simulated environment, we cannot talk about ego-centric views because the starting point is consistently an allocentric spatial representation from above, which does not align with the embodied perspective of the robot. While it is true that the generated texts could follow either an allocentric or egocentric description of the scene, the description cannot be equally generated if the initial spatial representation is from an overhead view versus a first-person perspective. Hence the evaluation of the model with the simulator is not perfectly scalable to the robot. For the embodied robot, the spatial arrangement is not as clearly defined as in the simulator, and factors such as people obstructing the robot's view or the robot moving can introduce confusion, which is not due to sensors or software but rather the inherent limitations of the egocentric perspective. I believe it is important for the authors to address and comment on this aspect, especially if they intend to propose a new approach to the social robotics community.
7. The middle part of the conclusion would be more appropriate as a continuation of the discussion, especially if it includes an additional analysis of the triangle inequality in the embedding space. Also, this aspect should be more thoroughly explored, as the current treatment feels too brief. It only arrives at the conclusion, but it seems a core component of the approach proposed.
8. In general, it would be helpful to more clearly distinguish between the methodology, results, discussion, and conclusion sections. While I understand that this is primarily a proof-of-concept paper, restructuring these sections would greatly improve the clarity of

the paper. A more defined separation will allow readers to better follow the logical flow of the work and enhance the overall readability. This restructuring would strengthen the paper by providing a clearer organization of the content.

Data availability: I tried to access the data at https://doi.org/10.5281/zenodo.10623138 . It does not seem the dataset contains all the data used in the analysis. If so, are the authors planning to release the entire dataset after approval?

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: human robot interaction, social robotics, cognitive robotics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 19 December 2024

https://doi.org/10.21956/openreseurope.18694.r47034

?

Bing Li
SCALab (UMR 9193), CNRS University of Lille, France, France, France
Tatjana A. Nazir

University Lille, Lille, France

Your Report

The study *"Social Embeddings: Concept and Initial Investigation"* introduces an innovative method to represent social situations as vector embeddings, enabling quantitative comparisons of their similarity. This process involves transforming social scenes into textual descriptions and converting them into embeddings using pre-trained large language models (LLMs). By operationalizing social similarity, the study provides a pseudo-metric for interpreting and comparing social dynamics, allowing robots to recognize patterns and adapt behavior in new contexts based on past experiences. While primarily aimed at advancing socially aware robot behavior, the approach also holds potential for interdisciplinary applications, including digital humanities. Serving as a foundational exploration, the study lays the groundwork for future research in enhancing robots' comprehension and responsiveness to complex social environments.

Major comments:

The concept of social embeddings provides a compelling framework for advancing HRI by quantifying and comparing social scenarios. However, the process of creating textual descriptions from real-world sensory data (e.g., video, audio) is underdeveloped. Real-world social interactions are rich with cues such as facial expressions, tone of voice, and gaze that are challenging to encode in language but critical for understanding social situations. Algorithm 1 risks oversimplifying social semantics as the complexity of real-world cues increases. Note that one of the strengths of text embeddings is that they flexibly map semantic properties and relations onto latent dimensions. However, when the authors used algorithm 1 to transform a social scenario into a description, they somewhat *mechanically* reduced the semantic properties and relations in the text. With this algorithm, people can debate about the necessity of using embedding. Because they can represent social scenes with diagonal matrices with social agents as column and row labels, and then calculate the similarities among the matrices. To address this issue, the authors should further discuss and explain how the complexity of algorithm 1 grows with more social cues coming into the scene. For example, in addition to "walking towards/away", "looking at", "talking to", what often happens in real social scene is that someone could "look up at", "look down at", "smile at", or "talk to someone with a smile" or even "talking to someone while smiling at another", etc. In short, the authors should tell readers how the complexity of algorithm 1 grows with the complexity of the social scene, which will likely grow indefinitely. This is especially important if this study targets improving real human-robot interactions.

Minor comments:

1. The results in Section 3.2, which investigate the invariance of embeddings to pragmatics, are insightful but lack clarity in their presentation. The use of cosine distances and edit (Levenshtein) distances provides an interesting comparison, but the visualization could be improved (e.g. the diagram below). Also, in order to draw conclusions on whether *embeddings do encode semantics of the scene description independently of the exact pragmatics of the description*, a statistical test showing the "self-other" fitting and the "self-self'" are different should be done. https://s3-eu-west-1.amazonaws.com/openreseurope/supplementary/17296/dcc4cd08-7734-454e-a322-0a807ce1ec48.pdf
2. Regarding the same social scene, how do the sentences "a faces b, b faces c" & "b faces c, a faces b" are distanced from each other in terms of edit/Levenshtein and cosine distance?
3. Can continuous distances be used in algorithm 1, instead of discrete categories "close", "medium", "far", in order to test the continuity property of the framework?

4. Since “it appears that egocentric descriptions yield slightly better social representations”, can we use angle degrees, instead of meters, as units of the distances between social agents?
5. The methodology section sometimes jumps into technical details without sufficiently framing the necessity and context of the approach.
6. Figures and tables, while informative, are not always clearly integrated into the text.
7. The introduction and discussion lacks contextualization within broader fields such as social psychology or HRI. It also does not sufficiently acknowledge limitations, such as the inability to capture nonverbal social cues.
8. The reference list is limited (17 sources) and misses key works in related fields such as embeddings-based approaches in other domains.
9. Lastly, the manuscript would benefit from a clearer articulation of its purpose—whether as a conceptual framework, a technical proof of concept, or an evaluation.

Conclusion

The concept of social embeddings offers an exciting pathway for robots to interpret human social environments. However, potential difficulties for the transition from proof of concept to practical utility should be addressed and a stronger statistical and theoretical framework and deeper engagement with prior research would improve this manuscript.

Is the work clearly and accurately presented and does it cite the current literature?

No

Is the study design appropriate and does the work have academic merit?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Not applicable

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Cognitive Sciences

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.