Predicting Human Interactivity State from Surrounding Social Signals

Youssef Mohamed
os19105@bristol.ac.uk
University of the West of England
Bristol Robotics Laboratory
Bristol, UK

Séverin Lemaignan
severin.lemaignan@brl.ac.uk
University of the West of England
Bristol Robotics Laboratory
Bristol, UK

ABSTRACT

This article presents the use of a multi-layered perceptron neural network to predict if one person in a group is being interactive or not, based on the social signals of the other group members. Interactivity state (as manually annotated post-hoc) was correctly predicted with 60% accuracy when using the person's own social signals (*self state*), but showed higher accuracy of 65% when using instead social signals from the surrounding group members, excluding the target person (*group members state*). These results are preliminary due to the limits of our dataset (a micro dataset of 6 participants – of which 3 are in frame – playing the social game *mafia*, with 734 frames). A post-hoc factor analysis reveals that facial actions units and the distance between the target person and the group members are the key features to consider when estimating interactivity state from surrounding social peers.

CCS CONCEPTS

• General and reference → Surveys and overviews; **• Human-centered computing → Interaction design**; **• Computing methodologies → Artificial intelligence**.

KEYWORDS

Human-robot interaction, Social Robots, Social Signal Processing

ACM Reference Format:
Youssef Mohamed and Séverin Lemaignan. 2021. Predicting Human Interactivity State from Surrounding Social Signals. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21 Companion), March 8–11, 2021, Boulder, CO, USA.* ACM, NY, NY USA, 4 pages. https://doi.org/10.1145/3434074.3447230

1 INTRODUCTION

Currently, in the psychology literature, much research is being carried out to analyze the internal states of humans and understand how emotions develop and affect behaviour [2, 6]. However, understanding internal states can be a tricky task, as robots have to use non-invasive and observational techniques to understand processes as complicated as emotions. Therefore, social psychology can be used as a stepping stone for further research, as it is easier to understand human behaviour as a collective group of behaviours and not on an individual level. As in each social situation, each person has a significant influence on the other [19]. Although the way peer influence is transmitted is still unclear in the psychology literature [13], the effects are observable and have been discussed extensively [1, 13–15]. Therefore, to prove the possibility of predicting behaviour based on peer influence, an MLP (Multilayer Perceptron) model is trained to predict if one person is being interactive in the situation or not (refer to Table 1).

2 METHODS

In this section several aspects of the methods used are going to be discussed: the creation and annotation of the data set, data extraction and prepossessing and the creation of the model and the process of features selection. The raw CSV file and the data processing script are open-source and can be downloaded from github.com/youssef266/MLP_Mafia.

2.1 Mafia Micro Data Set

Recording human-human and human-robot interactions have been used extensively as a tool to have better understanding of human behaviour and reactions in certain situations [5, 8–10]. However, using role playing games has been one the aspects that are less explored in the literature. Role playing can be used to provoke certain behaviours in a controlled environment [7], as it is the case in the game of MAFIA, which provokes the players to suspect and deceive each other while having a conversation on which player to eliminate from the game [3]. This kind of highly dynamic environment makes the game of mafia a perfect approach for analyzing behaviours, as it has been used by [4] to detect the deceptive roles, using only non-verbal audio cues.

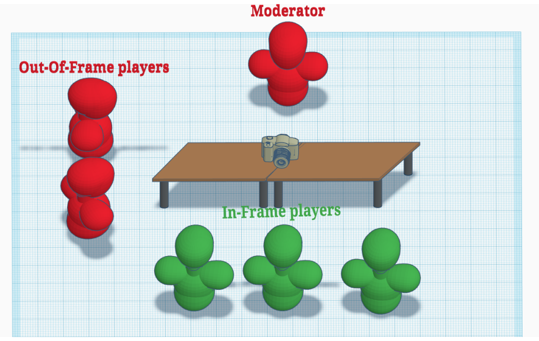
Therefore, the mafia micro data set was created, which included six people playing the game of MAFIA, only 3 of which are in the frame, as seen in Figure 1a.

Figure 1b shows a snapshot of the data set in which two of the participants are interacting with a person outside of the frame.

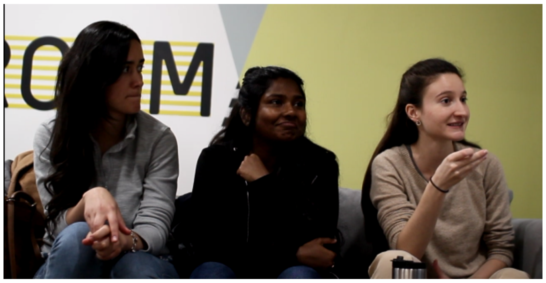
2.1.1 Annotation of interactivity state. To be able to annotate the video on how interactive the group members are within each of the interactions that occurred, an annotation schema was developed to assess the social presence of the players within the bounds of the interactions. Social presence relates directly to the inter-activeness of the players [12]. As defined by Saran et al. [18], social presence is the "degree of salience of the other person in a mediated communication

HRI '21 Companion, March 8–11, 2021, Boulder, CO, USA

ACM ISBN 978-1-4503-8290-8/21/03…$15.00
https://doi.org/10.1145/3434074.3447230



(a) Players in and out of frame.



(b) A snapshot from the mafia data set, showing the behaviour of the three participants. While the two rightmost participants are clearly engaged in a social interaction, the left participant seem less socially active at the time of this snapshot.

Figure 1: Game formation

and the consequent salience of their interpersonal interactions". Based on literature on the analysis of the social presence of groups of students and the level of affective behaviours shown [12, 16, 17], we annotated the three following salient features of interactiveness, and a participant would be marked as interactive whenever two of these three features would co-occur.

- Shared gaze
- Facially expressive
- Expressive body language

To be able to annotate the videos quickly and efficiently Elan[1] was used for segmentation and annotation. The annotation was made for each person separately, as it was more efficient to separate the people in each of the frames to make it easier for the machine learning model to use.

2.2 Data Extraction & Processing

The data has been extracted from the Mafia data set, using a ROS synchronizer[2] to filter and sync the published messages with the same timestamp in one CSV file. The data contained 734 instances, 447 of which the target person was not interactive. The features selected are shown in table 1. The table contains a range of features; each has a significant impact on human interactions. The features have been selected based on the available state of the art social extraction tools which include: distances with respect to each of the faces, the rotation with respect to each of the faces, Action units for each of the faces, upper body pose and the annotated interactivity state of each person.

Light pre-processing was performed (to eg compute angles and distances between participants based on their Cartesian coordinates), then split into 5 folds for cross-validation to insure that the model is not over-fitting.

The data was split into two subsets:

- Self state data: only the social signals related to the target person
- Group members state data: included the data of the surrounding group members, excluding the target person's social signals.

2.3 Model Creation & Feature Selection

An MLP neural network has been used to create a model that is able to classify if the participants are interactive or non-interactive in the social situation. The architecture of the self state subset model consists of 3 hidden layers that have 100,100 and 10 nodes respectively. On the other hand, for the group state subset model, only one hidden layer was used with 75 nodes. Both model's hyper-parameters were calculated using a grid search algorithm.

Furthermore, forward sequential feature selection (SFS) algorithm is a wrapper method that has been used to evaluate the attribute's relevance to the model's performance in relation to the accuracy. The SFS algorithm uses several greedy search methods to be able to reduce the dimensions of the feature vector used, which selects the subset that yields the highest accuracy of the model.

All data processing was performed using Python's `scikit-learn`[3] and `MLxtend`[4] libraries.

3 RESULTS

The results shown below are with respect to classifying the interactivity status of the three participants in the frame (P0, P1 and P2). The MLP model has been trained using either the features (Table 1) and annotated interactivity state of P0, P1 and P2, or the features of the two other participants only. The model's precision, recall, F1-score and ROC values were calculated to evaluate the performance of the model being trained on both the self state and the group members state data subsets.

In addition, the sequential feature selector's selected features for each person in the frame were also used to evaluate the model's behaviour and detect if it would be able to predict the interactivity status with a similar accuracy of the full data.

3.1 Model Performance

3.1.1 *Self State.*

The self state subset contained exclusively the data of the target person, which showed a an average of 60% accuracy across the 5 folds in predicting the target person. As seen in table 2, the average ROC value across the 5 folds is 0.48, which is a strong indicator that the model performs worst than chance.

[1] https://archive.mpi.nl/tla/elan
[2] http://wiki.ros.org/message_filters
[3] https://scikit-learn.org
[4] http://rasbt.github.io/mlxtend/

Table 1: Features used in training data.

Feature	Type	Values	Possible inference
Head distance with respect to other heads in real-world units (meters)	Float	Ranges from -1.7 to 1.8	Gaze, Proximity, Attention,
Head angle with respect to other heads in radians.	Float	Ranges from -0.7 to 0.9	Attention
The presence of Action Units of each of the heads.	Binary	0/1	Facial expression, Eye status (closed/Open), Level of facial expressiveness
Estimated emotion from facial expression	Nominal	Happy, Sad, Surprise, Anger, Neutral	Basic facial emotion, Facial expressiveness
Upper body pose of each person	Nominal	handsonface, armscrossed, handsraised, other	Body language, Basic engagement level, Body expressiveness level
The annotated interactivity status for each person in the frame	Binary	0/1	Expressive behavior as a human would see it.

Table 2: Self state subset scores

	Accuracy	Precision	Recall	F1-Score	ROC
(+/- stdv)	0.60 (+/- 0.01)	0.38 (+/- 0.39)	0.06 (+/- 0.08)	0.10 (+/- 0.13)	0.48 (+/- 0.16)

3.1.2 Group Members State. The group members data subset contained only the data of the surrounding group members, which excludes the target person's data ($((P0|P1, P2){+}(P1|P0, P2){+}(P2|P0, P1))$). The model has an accuracy of 65% predicting the target person's interactivity. Furthermore, the ROC value has an average of 0.65 with a standard deviation of 0.25 across the 5 folds.

Table 3: Group members state subset scores

Fold	Accuracy	Precision	Recall	F1-Score	ROC
(+/- stdv)	0.65 (+/- 0.19)	0.54 (+/- 0.26)	0.53 (+/- 0.33)	0.53 (+/- 0.29)	0.65 (+/- 0.25)

3.2 Sequential Feature Selector

3.2.1 Self State Subset. As shown in figure 2, the model needed only two features to get to 60%. Yet, the value did not change when features were being added. That would be an indication that the model is performing based mainly on chance and that the correlation between the features and the interactivity status is weak.

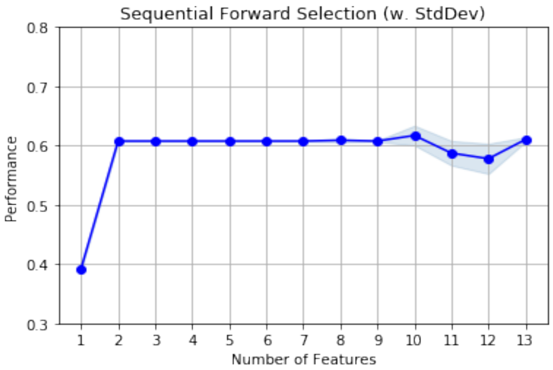


Figure 2: Accuracy after each feature in the self state subset

3.2.2 Group State Subset. The SFS algorithm has calculated the most essential 20 features out of the 46 used. The first 17 features were only the action units of both people in the frame, and then the last three were the interactivity status and the distance between both people and the target person. In figure 3 it is shown that the model performed best with only 17 features, to reach the accuracy of 70%, the features included the interactivity status of one the people in the frame, the action units of both people and the distance between the target person and one of the people in the frame.

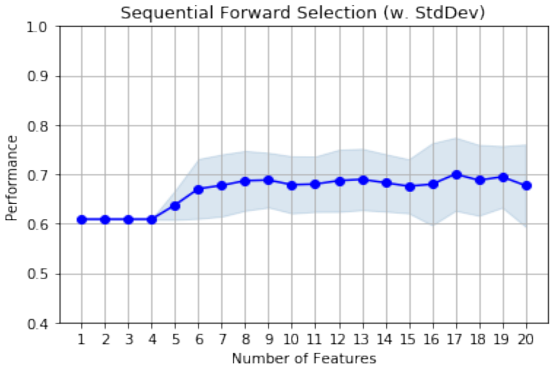


Figure 3: Accuracy after each feature in the group state subset

4 DISCUSSION

It is clear to see by comparing both model performances that using only the target person's data did not perform much better than chance. On the other hand, the model using the surrounding group members showed a higher potential of being able to predict the interactivity state of a person, it can be seen in table 3 that the accuracy and the ROC value was more than 0.5, which indicates that the model is able to perform better than chance. Hence, using more data can create a more robust, stable and more accurate model in the case of the group members state subset. Furthermore, the SFS algorithm showed the most essential features that can be used to produce a model with similar accuracy to the model that uses the whole data, the main features were the action units of the surrounding members in addition to their position relative to the target person.

5 FUTURE WORK

The data collected in this case is considered sparse as there was only 734 complete instances, with no missing values. The missing values were caused by the incapability of the used tools to detect the faces of the participants. Therefore, more computational power could have been used to increase the performance of the tools. In addition, a bigger data set could have been recorded to increase the number of detections with no missing values. The temporal aspect has not been considered in this approach as the problem has been simplified to being a matter of classification. Therefore, using a recurrent neural network (RNN) can be a possible solution for considering the temporal aspect. RNNs are usually used in natural language processing, nonetheless, in [11] it has been used for human action recognition. In addition, another study used a multi-component CNN-RNN approach to recognize basic emotions based on facial expressions. Recently in [20] an approach has been proposed to consider the spatio-temporal aspect of human behaviour recognition, the approach introduced a temporal layer in a CNN model and it preformed better than other models. In future works, a similar model would be incorporated to account for the temporal aspect, which was not considered using a simple MLP model.

6 CONCLUSION

The paper presented a model that is able to predict the interactivity status of a target person based only on the social signals of other people that are participating in the same social situation. Yet, the predictions were made on only 734 instances in a 20-minute video of 3 people playing the game MAFIA. As a result, this paper only discusses the potential of using this approach and the necessity that the model would be tested on larger data sets, to guarantee the robustness and stability of the model.

REFERENCES

[1] Gwen Rae Bachmann, Deborah Roedder John, and Akshay R Rao. 1993. Children's susceptibility to peer group purchase influence: an exploratory investigation. *ACR North American Advances* (1993).

[2] M. Bartlett, C. E. R. Edmunds, T. Belpaeme, S. Thill, and S. Lemaignan. 2019. What Can You See? Identifying Cues on Internal States from the Kinematics of Natural Social Interactions. *Frontiers in AI and Robotics* (2019). https://doi.org/10.3389/frobt.2019.00049

[3] Mark Braverman, Omid Etesami, and Elchanan Mossel. 2008. Mafia: A theoretical study of players and coalitions in a partial information environment. *Ann. Appl. Probab.* 18, 3 (06 2008), 825–846. https://doi.org/10.1214/07-AAP456

[4] G. Chittaranjan and H. Hung. 2010. Are you Awerewolf? Detecting deceptive roles and outcomes in a conversational role-playing game. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. 5334–5337. https://doi.org/10.1109/ICASSP.2010.5494961

[5] Claudio Coppola, Serhan Cosar, Diego R Faria, and Nicola Bellotto. 2017. Automatic detection of human interactions from rgb-d data for social activity classification. In *2017 26th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE, 871–876.

[6] Antonio R. Damasio, Thomas J. Grabowski, Antoine Bechara, Hanna Damasio, Laura L.B. Ponto, Josef Parvizi, and Richard D. Hichwa. 2000. Subcortical and cortical brain activity during the feeling of self-generated emotions. *Nature Neuroscience* 3, 10 (oct 2000), 1049–1056. https://doi.org/10.1038/79871

[7] John Derek Greenwood. 1983. Role-playing as an experimental strategy in social psychology. *European Journal of Social Psychology* 13, 3 (jul 1983), 235–254. https://doi.org/10.1002/ejsp.2420130304

[8] D. B. Jayagopi, S. Sheiki, D. Klotz, J. Wienke, J. Odobez, S. Wrede, V. Khalidov, L. Nyugen, B. Wrede, and D. Gatica-Perez. 2013. The vernissage corpus: A conversational Human-Robot-Interaction dataset. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 149–150. https://doi.org/10.1109/HRI.2013.6483545

[9] Iulia Lefter, Gertjan J Burghouts, and Leon JM Rothkrantz. 2014. An audio-visual dataset of human–human interactions in stressful situations. *Journal on Multimodal User Interfaces* 8, 1 (2014), 29–41.

[10] Séverin Lemaignan, Charlotte ER Edmunds, Emmanuel Senft, and Tony Belpaeme. 2018. The PInSoRo dataset: Supporting the data-driven study of child-child and child-robot social dynamics. *PloS one* 13, 10 (2018), e0205999.

[11] Wenbo Li, Longyin Wen, Ming-Ching Chang, Ser Nam Lim, and Siwei Lyu. 2017. Adaptive RNN Tree for Large-Scale Human Action Recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[12] Patrick R Lowenthal. 2009. Social presence. In *Encyclopedia of Distance Learning, Second Edition*. IGI Global, 1900–1906.

[13] Kim CIM Megens and Frank M Weerman. 2012. The social transmission of delinquency: Effects of peer attitudes and behavior revisited. *Journal of Research in Crime and Delinquency* 49, 3 (2012), 420–443.

[14] Sidharth Muralidharan and Linjuan Rita Men. 2015. How peer communication and engagement motivations influence social media shopping behavior: Evidence from China and the United States. *Cyberpsychology, Behavior, and Social Networking* 18, 10 (2015), 595–601.

[15] Francois Poulin, Thomas J Dishion, and Eric Haas. 1999. The peer influence paradox: Friendship quality and deviancy training within male adolescent friendships. *Merrill-Palmer Quarterly (1982-)* (1999), 42–61.

[16] Ruth Rettie. 2003. Connectedness, awareness and social presence. (2003).

[17] Liam Rourke, Terry Anderson, D Randy Garrison, and Walter Archer. 1999. Assessing social presence in asynchronous text-based computer conferencing. *The Journal of Distance Education/Revue de l'ducation Distance* 14, 2 (1999), 50–71.

[18] John Short, Ederyn Williams, and Bruce Christie. 1976. *The social psychology of telecommunications*. John Wiley & Sons.

[19] Kathryn A Urberg. 1992. Locus of peer influence: Social crowd and best friend. *Journal of Youth and Adolescence* 21, 4 (1992), 439–450.

[20] Fuguang Yao. 2020. Deep learning analysis of human behaviour recognition based on convolutional neural network analysis. *Behaviour & Information Technology* (2020), 1–9.