How Can I Help You? Comparing Engagement Classification Strategies for a Robot Bartender

Mary Ellen Foster
School of Mathematical and Computer Sciences
Heriot-Watt University, Edinburgh, UK
M.E.Foster@hw.ac.uk

Andre Gaschler and Manuel Giuliani
fortiss An-Institut der TU München
Munich, Germany
{gaschler,giuliani}@fortiss.org

ABSTRACT

A robot agent existing in the physical world must be able to understand the social states of the human users it interacts with in order to respond appropriately. We compared two implemented methods for estimating the engagement state of customers for a robot bartender based on low-level sensor data: a rule-based version derived from the analysis of human behaviour in real bars, and a trained version using supervised learning on a labelled multimodal corpus. We first compared the two implementations using cross-validation on real sensor data and found that nearly all classifier types significantly outperformed the rule-based classifier. We also carried out feature selection to see which sensor features were the most informative for the classification task, and found that the position of the head and hands were relevant, but that the torso orientation was not. Finally, we performed a user study comparing the ability of the two classifiers to detect the intended user engagement of actual customers of the robot bartender; this study found that the trained classifier was faster at detecting initial intended user engagement, but that the rule-based classifier was more stable.

Categories and Subject Descriptors: H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems – Evaluation/methodology; I.2.6 [Artificial intelligence]: Learning

Keywords: Social signal processing; supervised learning

A customer attracts the bartender's attention	
ROBOT:	[Looks at Customer 1] How can I help you?
CUSTOMER 1:	A pint of cider, please.
Another customer attracts the bartender's attention	
ROBOT:	[Looks at Customer 2] One moment, please.
ROBOT:	[Serves Customer 1]
ROBOT:	[Looks at Customer 2] Thanks for waiting. How can I help you?
CUSTOMER 2:	I'd like a pint of beer.
ROBOT:	[Serves Customer 2]

Figure 1: A socially aware robot bartender

1. INTRODUCTION

As robots become integrated into daily life, they must be able to deal with situations in which socially appropriate interaction is vital. In such a setting, it is not enough for a robot simply to achieve its task-based goals; instead, it must also be able to satisfy the social goals and obligations that arise through interactions with people in real-world settings. As a result, a robot requires not only the necessary physical skills to perform objective tasks in the world, but also the appropriate social skills to understand and respond to the intentions, desires, and affective states of the people it interacts with.

Building a robot to meet these social interaction goals presents state-of-the-art challenges to all components of an artificial system. In this paper, we focus on the task of input processing: the robot must be able to recognise and interpret multimodal social signals from its human partners (e.g., gaze, facial expression, and language) in order to respond appropriately. However, a state-of-the-art input processing component such as vision or speech recognition produces a continuous stream of noisy sensor data. In order for this information to be useful in an interactive system, all of this continuous, noisy, single-channel information must be combined into a discrete, cross-modal representation to allow the decision-making components to select appropriate behaviour. This is the task of *social signal processing*, a topic that has received increasing attention in recent years—e.g., see [33] for a recent survey.

This work takes place in the context of a socially aware robot bartender (Figure 1). The hardware for the robot bartender consists of two manipulator arms with grippers, mounted to resemble human arms. Sitting on the main robot torso is an animatronic talking head capable of producing facial expressions, rigid head motion, and lip-synchronised synthesised speech. The robot bartender supports interactions like the one shown in the figure, in which two customers enter the bar area and each attempt to order a drink from the bartender. Note that when the second customer appears while the bartender is engaged with the first customer, the bartender reacts by telling the second customer to wait, finishing the transaction with the first customer, and then serving the second customer. In the context of this initial bartending scenario, the main role of social signal processing is to estimate *desired engagement*: using the low-level sensor data to determine, for each customer in the scene, whether that customer currently requires attention from the system.

Bohus and Horvitz [5, 6] pioneered the use of data-driven methods for the task of engagement classification. They trained models designed to predict user engagement based on information from face tracking, pose estimation, person tracking, group inference, along with recognised speech and touch-screen events. After training, their model was able to predict intended engagement 3–4 seconds in advance, with a false-positive rate of under 3%. A number of more recent systems have also used machine learning to address this task. For example, Li et al. [24] estimated the attentional state of users of a robot in a public space, combining person tracking, facial expression recognition, and speaking recognition; the classifier performed well in informal real-world experiments. Castellano et al. [7] trained a range of classifiers on labelled data extracted from the logs of children interacting with a chess-playing robot, where the label indicated either high engagement or low engagement. They found that a combination of game context-based and turn-based features could be used to predict user level engagement with an overall accuracy of approximately 80%. McColl and Nejat [27] automatically classified the social accessibility of people interacting with their robot based on their body pose, with four possible levels of accessibility: the levels estimated by their classifier agreed 86% of the time with those of an expert coder. MacHardy et al. [26] classified the engagement states of audience members for an online lecture based on information from facial feature detectors; the overall performance was around on this binary classification task 72%.

Like the above systems, we also make use of data-driven methods for estimating the desired engagement of customers of the robot bartender. We begin with a simple, hand-coded, rule-based classifier based on the observation of human behaviour in real bars. Using an annotated corpus based on the sensor data gathered from an initial human-robot experiment, we then train a range of supervised-learning classifiers and compare them through cross-validation. The rule-based classifier and the top-performing trained classifier are then integrated into the full robot bartender system and compared experimentally through interactions with real human users.

2. SOCIAL STATE RECOGNITION IN THE ROBOT BARTENDER

The robot bartender incorporates a large number of hardware and software components; details of the architecture and components are presented in [10, 14]. The current work takes place in the context of the **Social State Recogniser** (SSR), whose primary role is to turn the continuous stream of sensor messages produced by the low-level input-processing components into a discrete representation of the world, the robot, and all entities in the scene, integrating social, interaction-based, and task-based properties; see [30] for a formal description of the inputs and outputs of the SSR. The SSR constantly monitors the state, and publishes a state-update event to the interaction manager every time there is a change which might require a response from the system. In addition to storing and discretising all of the low-level sensor information, the state manager also infers additional relations that are not directly reported by the sensors. For example, it fuses information from vision and speech to determine which user should be assigned a recognised spoken contribution, and estimates which customers are in a group. Most importantly in the current scenario—where one of the main tasks is to manage the engagement of multiple simultaneous customers, as in Figure 1—the SSR also informs the interaction manager every time a customer is seeking to engage.

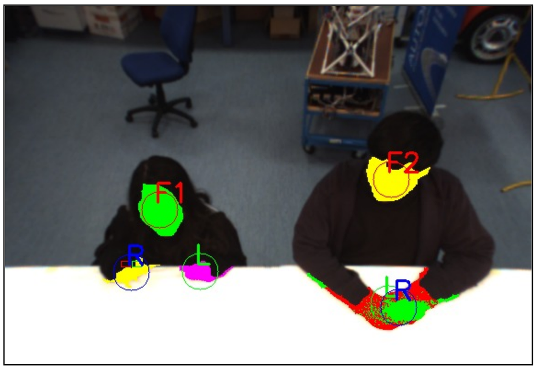

Figure 2: Output of face and hand tracking (image from [10])

To classify desired user engagement, the SSR makes use of low-level sensor data published on two input channels. The computer vision system [4, 29] tracks the location, facial expressions, gaze behaviour, and body language of all people in the scene in real time, using a set of visual sensors including two calibrated stereo cameras and a Microsoft Kinect [28] depth sensor. The data from the vision system is published as frame-by-frame updates multiple times a second. The other primary input modality in the system is linguistic [31], combining a speech recogniser with a natural-language parser to create symbolic representations of the speech from all users. For speech recognition, we use the Microsoft Speech API together with a Kinect directional microphone array; incremental hypotheses are published constantly, and recognised speech with a confidence above a defined threshold is parsed using a grammar implemented in OpenCCG [36] to extract the syntactic and semantic information.

Concretely, for this initial experiment in engagement classification, we make use of the following data from the input sensors:

- The (x, y, z) coordinates of each customer's head, left hand, right hand as reported by the vision system (Figure 2);
- The angle of each customer's torso in degrees, where 0° indicates that the customer is facing directly forwards; and
- An estimate of whether each customer is currently speaking, derived from the estimated source angle of each speech hypothesis along with the location information from vision.

We have implemented two strategies for estimating desired customer engagement using the above sensor properties: a classifier that uses a simple, hand-crafted rule based on the observation of natural interactions in a real bar, and a set of trained classifiers developed using supervised learning on an annotated human-robot corpus.

The **rule-based** engagement classifier relies on the signals observed in real bar customers who signalled that they wanted to engage with the bartender [19]: (1) standing close to the bar, and (2) turning to look at the bartender. These signals were extremely common in the natural data; and in a follow-up classification experiment based on still images and videos drawn from the natural data, they also proved both necessary and sufficient for detecting intended customer engagement [25]. Based on the details of the bartender environment, these signals were formalised as a rule-based classifier that defined a user to be seeking engagement exactly when (1) their head was less than 30cm from the bar, and (2) they were facing approximately forwards (absolute torso angle under 10°).

The **trained** classifiers, on the other hand, make use of a multimodal corpus derived from the system logs and annotated video recordings from the first user study of the robot bartender [10]. In

CVR Classifies using regression: the target class is binarised, and one regression model is built for each class value [11].
IB1 A nearest-neighbour classifier that uses normalised Euclidean distance to find the closest training instance [2].
J48 Classifies instances using a pruned C4.5 decision tree [32].
JRip Implements the RIPPER propositional rule learner [9].
LibSVM-0 Generates a Support Vector Machine using LIBSVM [8], with the default γ value of 0.
LibSVM-1 Uses LibSVM with $\gamma = 0.0001$.
Logistic Multinomial logistic regression with a ridge estimator [23].
NaiveBayes A Naïve Bayes classifier using estimator classes [20].
ZeroR Baseline classifier; always predicts the most frequent value.

Figure 3: Classifiers considered

particular, the engagement state of each customer visible in the scene was annotated with one of three levels: *NotSeekingEngagement*, *SeekingEngagement*, and *Engaged*. For the current classification task—where we aim to detect users who have not yet engaged with the system but are seeking to do so—the *Engaged* state is not relevant, so the corpus was based on the time spans annotated with one of the other labels. In total, the corpus consisted of 5090 instances: each instance corresponded to a single frame from the vision system, and contained the low-level sensor information for a single customer along with the annotated engagement label. 3972 instances were in the class *NotSeekingEngagement*, while 1118 were labelled as *SeekingEngagement*. Using the Weka data mining toolkit [16], we then trained a range of supervised-learning classifiers on this corpus, using a set of classifiers (Figure 3) designed to provide good coverage of different classification styles, based on those listed in the Weka primer [1]. Since the performance of the default (Radial Basis Function) kernel used by LIBSVM depends heavily on the value of the γ parameter, which controls the width of the kernel [18], we included two versions of this classifier: one using the default value of 0 (LibSVM-0), and one where γ was set to 0.0001 (LibSVM-1). All other classifiers were used in the default configuration as provided by Weka version 3.6.8.

3. OFFLINE EVALUATION

As a first step, we carried out an offline experiment to compare the performance of the trained classifiers with each other and with that of the rule-based classifier. This study provides an initial indication of which classification strategies are and are not suitable for the type of data included in the training corpus, and also gives an indication of the performance of the rule-based classifier on the same data.

3.1 Cross-validation

We compared the performance of all of the classifiers through 10-fold cross-validation on the training corpus. For each classifier, we computed the following measures: the overall classification accuracy, the area under the ROC curve (AUC), along with the weighted precision, recall, and F measure. Note that the baseline accuracy score for this binary classification task is the size of the larger class (*NotSeekingEngagement*): 3972/5090 = 0.78. The results of this evaluation are presented in Table 1, sorted by accuracy; the overall performance of the hand-coded rule on the full training corpus is also included. The groupings in Table 1 reflect differences among the accuracy scores that were significant at the $p < 0.01$ level on a paired T test based on 10 independent cross-validation runs. In other words, the IB1 classifier had the highest performance on this measure; the LibSVM-1, J48, CVR and JRip classifiers were statistically indistinguishable from each other; the LibSVM-0, Logistic, and ZeroR classifiers were again indistinguishable (these classifiers generally labelled all instances as *NotSeekingEngagement*); while the NaiveBayes classifier and the hand-coded rule had the lowest overall accuracy by a significant margin. Figure 4 shows the ROC curves for all classifiers based on the *SeekingEngagement* class: as expected, the curves for all of the high-performing classifiers are close to optimal, while those for the other classifiers are closer to the chance performance of the baseline ZeroR classifier.

Classifier	Accuracy	AUC	Precision	Recall	F
IB1	0.960	0.932	0.957	0.958	0.957
LibSVM-1	0.931	0.871	0.931	0.932	0.930
J48	0.924	0.919	0.925	0.925	0.925
CVR	0.921	0.960	0.911	0.912	0.912
JRip	0.911	0.868	0.913	0.914	0.913
LibSVM-0	0.790	0.521	0.830	0.790	0.706
Logistic	0.780	0.739	0.727	0.781	0.710
ZeroR	0.780	0.500	0.609	0.780	0.684
NaiveBayes	0.669	0.656	0.726	0.662	0.685
Hand-coded rule	*0.655*	*na*	*0.635*	*0.654*	*0.644*

Table 1: Cross-validation results, grouped by accuracy

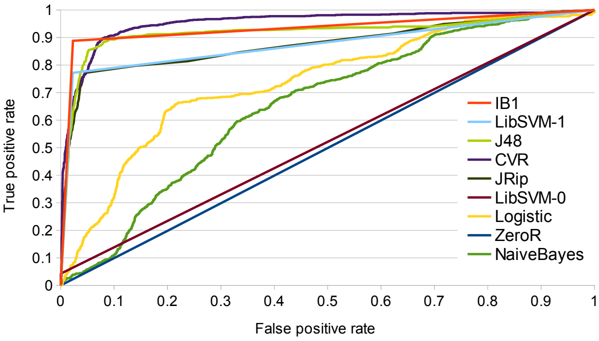

Figure 4: ROC curves for *SeekingEngagement* class

3.2 Attribute selection

The above cross-validation results made use of the full set of sensor attributes included in the corpus; however, it is likely that not all of the sensor data is equally informative for the classification task. To get a better assessment of which sensor data was most relevant, we carried out two forms of attribute selection. We first determined the sensor attributes that were the most informative for each of the individual classifiers, using a wrapper method [22] to explore the relationship between the algorithm and the training data. We then analysed the corpus as a whole using Correlation-Based Feature Selection (CBF) [17], a general-purpose selection method known to have good overall performance [15].

	Face			HandL			HandR			Ori	Spk	Acc
	x	y	z	x	y	z	x	y	z			
IB1	•	•	•	•			•	•	•		•	0.963
LibSVM-1	•	•	•								•	0.938
J48	•	•	•	•		•	•	•	•		•	0.932
CVR	•	•	•		•	•	•	•	•		•	0.926
JRip	•	•	•	•		•	•	•			•	0.921
LibSVM-0	•						•				•	0.830
Logistic												0.780
ZeroR												0.780
NaiveBayes			•	•	•						•	0.786
Hand-coded rule			•							•		*0.655*
CBF	•	•	•	•	•		•					

Table 2: Output of attribute selection

The results of this attribute selection process are shown in Table 2. The main body of the table indicates with a bullet (•) the attributes that were determined to be most informative for each of the classifiers; for reference, the last row shows the two features that were used by the rule-based classifier (z face location and body orientation). The final *Acc* column shows the cross-validation accuracy of a classifier making use only of the selected attributes. As can be seen, most of the high-performing classifiers made use of the full 3D location of the customer's head, along with the location of the hands and the "speaking" flag. The accuracy of most classifiers was very slightly better with the classifier-specific attribute subset when compared to the results from Table 1, but in no cases was this improvement statistically significant. The bottom row of the table shows the attributes that were found to be most informative by the CBF selector, which were similar to those used by the high-performing classifiers: namely, the full 3D position of the customer's head, along with some of the hand coordinates. The selected attributes correspond very well with prior results in our human-human interaction studies [12, 13].

It is notable that body orientation—which was one of the two main engagement-seeking signals found in the human-human data, and which was found to be necessary for making offline engagement judgements based on that same data—was not determined to be informative by any of the attribute selectors. This is most likely due to the performance of the initial vision system that was used to create the corpus data: the body orientation was often either incorrectly detected or not detected at all, making this attribute unreliable for engagement classification. The unreliability of this signal in the corpus data likely also affected the cross-validation performance of the hand-coded rule, which had lower accuracy than the baseline ZeroR classifier. Also, the right hand was generally found to be more informative than the left: this is probably because, assuming that most customers were right-handed, they would have used this hand more often, thus providing more useful vision data.

4. ONLINE EVALUATION

The offline results presented in the preceding section are promising: in cross-validation against real sensor data, the top-performing IB1 classifier correctly labelled over 96% of the instances. However, this study was based on frame-by-frame accuracy; and as Bohus and Horvitz [6] point out, for this sort of classifier, a better run-time evaluation is one that measures the errors per person, not per frame.

As a step towards such an evaluation, we therefore integrated the top-performing trained classifier into the robot bartender's Social State Recogniser and tested its performance against that of the rule-based classifier through an online evaluation, with human participants playing the role of customers for the robot bartender. This study used the drink-ordering scenario illustrated in Figure 1: two customers approached the bar together and attempted to engage with the bartender, and—if successful—each ordered a drink. The bartender was static until approached by a customer, and did not engage in any interaction other than that required for the target scenario.

The robot bartender used in this study was similar to the one used in the initial user evaluation [10], but with updates to all components: in particular, the vision system incorporated an improved method of estimating torso orientation [29]. For half of the trials, the SSR used the rule-based engagement classifier, while for the rest, it instead used the trained IB1 classifier. After each trial, the participants answered a short questionnaire regarding their experience of interacting with the bartender. In addition to the questionnaire, we also gathered a range of other measures assessing the performance of the two classifiers based on data gathered from the system log files.

4.1 Participants

41 participants (29 male), drawn from university departments outside the robotics group involved in developing the bartender, took part in this experiment. The mean age of the participants was 27.8 (range 16–50), and their mean self-rating of experience with human-robot interaction systems was 2.51 on a scale of 1–5. Participants were given the choice of carrying out the experiment in German or English; 27 chose to use German, while 14 chose English.

4.2 Scenario

The study took place in a lab, with lighting and background noise controlled as far as possible. In each trial, the participant approached the bartender together with a confederate, with both customers seeking to engage with the bartender and order a drink (as in Figure 1). Each participant was given a list of the possible drinks that could be ordered (Coke or lemonade), but was not given any further instructions. The robot was static until approached by a customer, and the confederate did not attempt to speak at the same time as the participant. After each interaction was completed, the participant completed a short computer-based questionnaire. Each participant carried out two interactions, with the order and selection of classifiers counter-balanced across participants.

4.3 Dependent measures

We gathered two classes of dependent measures: objective measures derived from the system logs, and subjective measures gathered from the questionnaire.

4.3.1 Objective measures

For this study, we computed several objective measures which specifically address the interactive performance of the two engagement classifiers. Note that the ground-truth data about the participants' actual behaviour is not yet available, as the videos from this study have not been annotated. However, based on the scenario (Figure 1), it is reasonably safe to assume that the majority of customers were seeking to engage with the bartender as soon as they appeared in the scene, and that the participants behaved similarly in the two classifier conditions. We collected the following objective measures:

Detection rate How many of the customers detected in the scene were classified as seeking to engage. Under the above assumptions, this measure assesses the accuracy of the two classifiers.

Initial detection time The average delay between a customer's initial appearance in the visual scene and the time that they were considered to be seeking engagement. Again, under the assumption that all participants behaved similarly, this measure assesses the relative responsiveness of the two engagement classifiers.

System response time The average delay between a customer's initial appearance in the visual scene and the time that the system generated a response to that customer. Since the system would only respond to customers that were detected as seeking engagement, this is a secondary measure of classifier responsiveness, but one that is more likely to have been noticed by the participants.

Drink serving time The average delay between a customer's initial appearance in the visual scene and the time that the system successfully served them a drink. Since serving a drink ultimately depends on successful engagement between the customer and the bartender, this is an even more indirect measure of responsiveness.

Number of engagement changes The average number of times that the classifier changed its estimate of a user's engagement-seeking state over the course of an entire experiment run. In the experimental scenario, only the initial detection affected the system behaviour: as soon as a customer was determined to be seeking engagement, the system would engage with them and the interaction would continue. However, the engagement classifier remained active throughout a trial, so this measure tracks the performance over time. Although the actual behaviour of the experimental participants is not known, we assume that it was similar across the two groups, so any difference on this measure indicates a difference between the classifiers.

4.3.2 Subjective measures

After each interaction, the participant filled out the short electronic questionnaire shown in Figure 5; a German translation was also available for participant doing the experiment in German.

1. Did you successfully order a drink from the bartender? [Y/N]

   Please state your opinion on the following statements:
   *[1:strongly disagree; 2:disagree; 3:slightly disagree;
   4:slightly agree; 5:agree; 6:strongly agree]*

2. It was easy to attract the bartender's attention [1-6]
3. The bartender understood me well [1-6]
4. The interaction with the bartender felt natural [1-6]
5. Overall, I was happy about the interaction [1-6]

Figure 5: Post-interaction questionnaire

4.4 Results

A total of 81 interactions were recorded in this study. However, due to technical issues with the system, only 58 interactions could be analysed, involving data from 37 of the 41 subjects: 26 interactions using the rule-based classifier, and 32 using the trained IB1 classifier. All results below are based on those 58 interactions.

4.4.1 Objective measures

Table 3 summarises the objective results, divided by the classifier type. Overall, the detection rate was very high, with 98% of all customers determined to be seeking engagement, generally within 4–5 seconds (and, in many cases, in under one second). The robot acknowledged a customer on average about 6–7 seconds after they first became visible, and a customer received a drink about a minute after their initial appearance—note that this last number includes the 20 seconds taken by the robot arm to physically grasp and hand over the drink. Over the course of an entire interaction, a customer's estimated engagement changed an average of about 13 times.

Each study participant took part in two interactions; however, as mentioned above, due to technical issues we could not analyse the full paired data. Instead, we analysed the data using a linear mixed model [3, 35], treating the participant identifier as a random factor, with the classification strategy and all demographic features included as fixed factors. This analysis found that the effect of the classification strategy on the number of changes in estimated engagement was significant at the $p < 0.05$ level; however, while the numbers in Table 3 suggest that the trained classifier was somewhat more responsive, none of those differences were found to be significant.

Several demographic factors also affected the objective results: the participants who carried out the experiment in German took significantly longer to receive their drinks than did those who interacted in English (48.1 vs. 62.0 seconds; $p < 0.05$), while the classifiers changed their estimate of the female participants' engagement state significantly more often over the course of an interaction (21.1 vs. 13.3 times; also $p < 0.05$).

Measure	Rule (sd)	Trained (sd)
Detection rate	0.98 (0.10)	0.98 (0.09)
Time to first detection	5.4 (7.9)	4.0 (9.7)
Time to system response	7.0 (7.9)	6.4 (10.4)
Time to drink served	62.2 (22.2)	53.7 (14.0)
Num. engagement changes	*12.0 (10.2)*	*17.6 (7.6)*

Table 3: Objective results (significant difference highlighted)

Question	Rule (sd)	Trained (sd)
Success	0.88 (0.32)	0.88 (0.33)
Attention	4.1 (1.6)	4.1 (1.8)
Understand	4.0 (1.8)	4.0 (1.8)
Natural	3.0 (1.3)	2.9 (1.6)
Overall	4.0 (1.6)	3.7 (1.7)

Table 4: Subjective results (all differences n.s.)

4.4.2 Subjective measures

The results from the subjective questionnaire are summarised in Table 4. In general, the participants gave the system reasonably high scores on perceived success, ease of attracting attention, understandability, and overall satisfaction, with somewhat lower scores for naturalness. The linear mixed model found that the choice of classifier had no significant effect on any of these measures. However, as in the preceding section, the demographic features of the participants also had a significant effect on the subjective results. There were two main effects in the models: the perceived success was significantly lower for participants with more knowledge of human-robot interaction ($R^2 = 0.13, p < 0.005$), while the participants who interacted in German gave significantly lower answers to the final, overall item on the questionnaire (3.5 vs. 4.5; $p < 0.05$).

We also carried out a PARADISE-style [34] stepwise regression to determine which objective features had the greatest effect on the participants' subjective responses. The details of this analysis are presented in [21]; in summary, all of the subjective responses were positively affected by the objective task success (i.e., the number of drinks served); the number of attempted user turns discarded due to low ASR confidence negatively affected most of the subjective questions; while various measures of dialogue efficiency (such as the system response time and the time taken to serve drinks) also had a significant impact, with longer interactions generally resulting in lower subjective scores.

4.5 Discussion

The objective results of this study indicate that the system was generally successful both at detecting customers who wanted to engage with it and at serving their drinks: despite the minimal instructions given to the participants, the objective success rate was very high. The choice between the two classification strategies had one main objective effect: the trained classifier changed its estimate of a customer's engagement state more frequently than did the rule-based classifier. While the trained classifier also appears to have been more responsive than the rule-based classifier, there was no significant difference found. The choice of classifier had no significant effect on any of the responses on the subjective questionnaire.

The demographics had several effects on the results. First, the participants who used German took significantly longer to receive their drink, and also gave lower overall ratings to the system. We suspect that this was likely due to the decreased performance of the Kinect German language model, which was added to the Speech API much more recently than the English recognition: on average, nearly twice as many attempted user turns were discarded due to low confidence for the German participants (4.1 per interaction) as for the English participants (2.2). Also, both classifiers' estimate of customer engagement changed more often over the course of a trial for the female participants than for the male participants: we hypothesise that this may be due to the vision system having been trained primarily on images of male customers. Finally, participants with more knowledge of human-robot interaction rated the perceived success significantly lower; note that perceived success was also significantly correlated with actual success as measured by number of drinks served ($R^2 = 0.176, p < 0.001$).

Note that all of the objective measures are based only on the data from the log files, along with some underlying assumptions about user behaviour based on the scenario given to the participants (Figure 1): namely, we assume that all customers were seeking to engage with the bartender from the moment that they appeared, and that the behaviour of the participants in the two conditions did not differ over the course of an interaction. The difference in classifier stability between male and female participants suggests that this assumption may not hold in practice; however, to assess the true performance of the classifiers, we require ground-truth data as to the actual engagement-seeking behaviour of the customers in the scene. Such ground-truth information will also allow us to analyse the impact of the demographic factors more directly. For this reason, we are currently annotating the video recordings from this study to add that information, and will carry out further analysis once the annotation is completed.

5. CONCLUSIONS AND FUTURE WORK

We have presented the role of social state recognition in the context of a socially aware robot bartender, and have described two approaches to the particular task of estimating customers' engagement intentions: the first version used a hand-coded rule based on findings from annotated human behaviour in real bars, while for the second version, we trained a range of supervised-learning classifiers using a multimodal corpus based on user interactions with the initial system. In a cross-validation study using real sensor data, nearly all of the classifiers significantly outperformed the hand-coded rule. The best-performing classifier based on accuracy was the instance-based IB1 classifier, which had an overall accuracy of 0.960 in frame-based cross-validation. When we carried out feature selection, it was found that the most informative features were the 3D position of the customer's head, along with some of the coordinates of their hands; body orientation—which was one of the two features used by the rule-based classifier—was actually not informative based on the corpus data, which we hypothesise was due to the noisiness of this signal in the vision data used for training.

In a user study comparing the rule-based classifier with the trained IB1 classifier in the context of the full robot bartender system, the trained classifier changed its estimate of the customers' engagement state significantly more often over the course of an interaction, suggesting that it is less stable; however, until the ground truth data is available of user behaviour, it is not clear which of the two classifiers actually performed better on this measure. The trained classifier also detected intended user engagement somewhat more quickly, leading to a mild (non-significant) improvement in system responsiveness. The choice of classifier did not have a significant impact on the users' subjective opinions of the robot bartender. Several demographic factors did have an impact on the study results, including the participants' gender and experience with human-robot interaction systems, along with the language in which they chose to interact.

This initial study has confirmed that, as in other similar domains, data-driven techniques are a suitable mechanism for social signal processing for the robot bartender. However, this study has several limitations. First, it addressed only a single, simple, binary classification task; also, it considered only a subset of the available properties from the input sensors, and did not make any use of the interaction history. Also, all of the participants in the user evaluation were instructed to seek to engage with the bartender from the start of the interaction, so we did not test the classifiers with any negative examples. Finally, the objective measures are based purely on the sensor estimates from the log files, rather than on the actual user behaviour: for example, we do not know either the actual engagement actions of each user or the actual speech that they used.

The immediate next task in this work is to annotate the user behaviour in the video recordings of the interactions from this study: this will allow more detailed objective measures to be gathered, and can also form the basis of a more sophisticated multimodal corpus incorporating state features such as the hypotheses from the speech recogniser and the history of the interaction, along with additional vision properties such as the customers' face orientations, facial expressions, and body gestures. The labels in this corpus will also incorporate richer high-level customer features such as group membership; new models based on these corpora will be trained and integrated into the system, and their performance will be assessed through further user evaluations. In future user studies, we will also take care to control demographic factors such as gender and language to ensure that the evaluation gives an assessment of classifier performance that is as accurate as possible.

6. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 270435, JAMES: Joint Action for Multimodal Embodied Social Systems (`james-project.eu`). Thanks to Ingmar Kessler for helping run the experiment.

7. REFERENCES

[1] Weka primer. `http://weka.wikispaces.com/Primer`.

[2] D. Aha and D. Kibler. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.

[3] R. Baayen, D. Davidson, and D. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412, 2008. doi:10.1016/j.jml.2007.12.005.

[4] H. Baltzakis, M. Pateraki, and P. Trahanias. Visual tracking of hands, faces and facial features of multiple persons. *Machine Vision and Applications*, 23(6):1141–1157, 2012. doi:10.1007/s00138-012-0409-5.

[5] D. Bohus and E. Horvitz. Dialog in the open world: platform and applications. In *Proceedings of ICMI-MLMI 2009*, pages 31–38, Cambridge, MA, Nov. 2009. doi:10.1145/1647314.1647323.

[6] D. Bohus and E. Horvitz. Learning to predict engagement with a spoken dialog system in open-world settings. In *Proceedings of SIGDIAL 2009*, pages 244–252, 2009.

[7] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. McOwan. Detecting engagement in HRI: An exploration of social and task-based context. In *Proceedings of SocialCom'12*, pages 421–428, Sept. 2012. doi:10.1109/SocialCom-PASSAT.2012.51.

[8] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3): 27:1–27:27, May 2011. doi:10.1145/1961189.1961199.

[9] W. W. Cohen. Fast effective rule induction. In *Twelfth International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann, 1995.

[10] M. E. Foster, A. Gaschler, M. Giuliani, A. Isard, M. Pateraki, and R. P. A. Petrick. Two people walk into a bar: Dynamic multi-party social interaction with a robot agent. In *Proceedings of ICMI 2012*, Oct. 2012.

[11] E. Frank, Y. Wang, S. Inglis, G. Holmes, and I. Witten. Using model trees for classification. *Machine Learning*, 32(1): 63–76, 1998.

[12] A. Gaschler, K. Huth, M. Giuliani, I. Kessler, J. de Ruiter, and A. Knoll. Modelling state of interaction from head poses for social Human-Robot Interaction. In *Proceedings of the Gaze in Human-Robot Interaction Workshop held at the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2012)*, Boston, MA, March 2012.

[13] A. Gaschler, S. Jentzsch, M. Giuliani, K. Huth, J. de Ruiter, and A. Knoll. Social Behavior Recognition using body posture and head pose for Human-Robot Interaction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, October 2012. doi:10.1109/IROS.2012.6385460.

[14] M. Giuliani, R. P. A. Petrick, M. E. Foster, A. Gaschler, A. Isard, M. Pateraki, and M. Sigalas. Comparing task-based and socially intelligent behaviour in a robot bartender. In *Proceedings of the 15th International Conference on Multimodal Interfaces (ICMI 2013)*, Sydney, Australia, Dec. 2013.

[15] M. Hall and G. Holmes. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering*, 15(6):1437–1447, 2003. doi:10.1109/TKDE.2003.1245283.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations Newsletter*, 11(1):10–18, Nov. 2009. doi:10.1145/1656274.1656278.

[17] M. A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)*, pages 359–366, 2000.

[18] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 15 April 2010. `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`.

[19] K. Huth, S. Loth, and J. De Ruiter. Insights from the bar: A model of interaction. In *Proceedings of Formal and Computational Approaches to Multimodal Communication*, Aug. 2012.

[20] G. H. John and P. Langley. Estimating continuous distributions in Bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Mateo, 1995.

[21] S. Keizer, M. E. Foster, O. Lemon, A. Gaschler, and M. Giuliani. Training and evaluation of an MDP model for social multi-user human-robot interaction. In *Proceedings of the 14th Annual SIGdial Meeting on Discourse and Dialogue*, 2013.

[22] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997.

[23] S. le Cessie and J. van Houwelingen. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201, 1992.

[24] L. Li, Q. Xu, and Y. K. Tan. Attention-based addressee selection for service and social robots to interact with multiple persons. In *Proceedings of the Workshop at SIGGRAPH Asia*, WASA '12, pages 131–136, 2012. doi:10.1145/2425296.2425319.

[25] S. Loth, K. Huth, and J. P. De Ruiter. Automatic detection of service initiation signals used in bars. *Frontiers in Psychology*, 4(557), 2013. doi:10.3389/fpsyg.2013.00557.

[26] Z. MacHardy, K. Syharath, and P. Dewan. Engagement analysis through computer vision. In *Proceedings of CollaborateCom 2012*, pages 535–539, Oct. 2012.

[27] D. McColl and G. Nejat. Affect detection from body language during social HRI. In *Proceedings of 2012 IEEE RO-MAN*, pages 1013–1018, Sept. 2012. doi:10.1109/ROMAN.2012.6343882.

[28] Microsoft Corporation. Kinect for Windows. URL `http://www.microsoft.com/en-us/kinectforwindows/`.

[29] M. Pateraki, M. Sigalas, G. Chliveros, and P. Trahanias. Visual human-robot communication in social settings. In *Proceedings of ICRA Workshop on Semantics, Identification and Control of Robot-Human-Environment Interaction*, 2013.

[30] R. P. A. Petrick and M. E. Foster. Planning for social interaction in a robot bartender domain. In *Proceedings of the ICAPS 2013 Special Track on Novel Applications*, Rome, Italy, June 2013.

[31] R. P. A. Petrick, M. E. Foster, and A. Isard. Social state recognition and knowledge-level planning for human-robot interaction in a bartender domain. In *AAAI 2012 Workshop on Grounding Language for Physical Systems*, Toronto, ON, Canada, July 2012.

[32] R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[33] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1): 69–87, Jan. 2012. doi:10.1109/T-AFFC.2011.27.

[34] M. Walker, C. Kamm, and D. Litman. Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6(3&4):363–377, 2000. doi:10.1017/S1351324900002503.

[35] B. West, K. B. Welch, and A. T. Galecki. *Linear mixed models: a practical guide using statistical software*. CRC Press, 2006.

[36] M. White. Efficient realization of coordinate structures in Combinatory Categorial Grammar. *Research on Language and Computation*, 4(1):39–75, 2006. doi:10.1007/s11168-006-9010-2.