Rapport Prediction Using Pairwise Learning in Dyadic Conversations Among Strangers and Among Friends

Takato Hayashi[1], Ryusei Kimura[1], Ryo Ishii[2], Fumio Nihei[2], Atsushi Fukayama[2], and Shogo Okada[1(✉)]

[1] Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, Japan
okada-s@jaist.ac.jp
[2] Human Informatics Laboratories, NTT Corporation, Yokosuka-shi, Kanagawa, Japan

Abstract. Automatic rapport prediction is a key component in the creation of socially aware conversational agents. In our study, we aim to automatically predict a speakers' subjective rapport using their nonverbal (acoustic and facial) cues during conversations. While cues indicating rapport vary according to social relations between speakers, few studies have investigated an effective modality or combination of modalities for predicting subjective rapport. To fill this research gap, we collected both first-meeting (FM) and friend (FR) conversations from the same participants. Then, we addressed predicting subjective rapport using a common framework in both FM and FR conversations. Predicting subjective rapport is often formulated as a regression task that directly predicts rapport ratings. However, regression is not a suitable approach because it does not consider individual differences and ambiguity in subjective ratings. Thus, we adopted pairwise learning (PL). PL overcomes individual differences and ambiguity in subjective ratings because PL does not directly use rapport ratings. Our experimental results showed that PL is a more appropriate approach than regression for predicting conversations in both FM and FR conversations. We also reported an effective modality or combination of modalities for predicting subjective rapport in FM and FR conversations, respectively.

Keywords: Rapport · Pairwise Learning · Affective Computing · Emotion · Nonverbal Communication

1 Introduction

Building rapport among speakers is essential for successful relations. This study aims to automatically predict the degree of subjective rapport using a speaker's nonverbal cue in a conversation. If rapport prediction is possible, subjective rapport can be recorded in conjunction with the content of the conversation for each speaker. This recorded information provides knowledge about the preferred conversation content of a speaker. This information, therefore, is useful to personalize a conversational agent to a specific speaker.

A. Coman and S. Vasilache (Eds.): HCII 2024, LNCS 14705, pp. 17–28, 2024.
https://doi.org/10.1007/978-3-031-61312-8_2

Beyond linguistic cues, rapport is conveyed through various nonverbal cues [1]. Therefore, researchers in affective computing have focused on automatically predicting rapport using verbal/nonverbal cues. Hagad et al. predicted rapport in dyadic conversations [2]. Müller et al. addressed detecting low rapport in group interactions [3]. Previous studies developed models to predict rapport in peer tutoring [4,5]. Madaio also indicated that cues for predicting rapport differ between peer tutoring among friends and among strangers [5]. Despite this finding, differences in effective cues for predicting rapport between natural conversations among friends and among strangers have not been explored.

To fill this research gap, we address predicting subjective rapport using a common framework in both natural conversations among friends and among strangers. The previous study finds that nonverbal cues indicating rapport vary according to social relations between speakers [1]. Predicting subjective rapport is often formulated as a regression task that directly predicts rapport ratings [6,7]. However, regression is not a suitable approach because it does not consider **individual differences** and **ambiguity** in subjective ratings. Thus, we adopt **Pairwise learning** (PL) to alleviate these problems.

First, subjective ratings have individual differences, which are caused by the *perceiver effect* [8] and the *response style* [9]. The *perceiver effect* is the tendency of perceivers to rate items for all targets in a particular way (e.g., positivity) [8]. For example, some perceivers often rate rapport for all targets positively, and others rate it negatively. The *response style* (RS) is a tendency of perceivers to rate items using specific categories regardless of content (e.g., extreme/midpoint RS) [10]. For example, perceivers with extreme/midpoint RS prefer the ends/center of the scale. Second, subjective affect ratings are ambiguous. When a perceiver is asked to rate the same item twice, their ratings are not necessarily the same [11]. Due to individual differences and ambiguity in subjective ratings, it is challenging for a regression model to learn the mapping from a perceiver's behavior to their rapport ratings.

PL is an attractive alternative approach. In PL, a model is trained to predict ordinal relations between two conversations based on rapport ratings reported by the same perceiver. PL overcomes individual differences and ambiguity in subjective ratings because PL does not directly use rapport ratings. Although there is enough evidence to show that PL has significant advantages over regression for emotion recognition [12,13], few studies have explored PL to predict subjective ratings in interpersonal perceptions (e.g., rapport) [7]. Hayashi et al. showed that PL is a more appropriate approach than regression for predicting subjective rapport in natural conversation among strangers [7].

Our study is composed of three main contributions. First, we collect online dyadic conversations in which the same participant communicates with multiple strangers and friends. Second, we investigate whether PL improves predictive performances of subjective rapport and whether it is superior to regression in not only first-meeting (FM) conversations but also friend (FR) conversations. We use three evaluation metrics to measure ranking performance because we address the task of ranking conversations according to the degree of subjective rapport. Third, we demonstrate an effective modality or combination of modalities for

Table 1. Dataset Summary.

First-meeting Conversation	
No. of participants (male)	69 (35)
No. of pairs of participants	96
No. of conversations	288
Friend Conversation	
No. of participants (male)	32 (16)
No. of pairs of participants	48
No. of conversations	144

predicting subjective rapport in FM and FR conversations, respectively. In our experiments, we use acoustic and facial features.

2 Data

We collected online dyadic conversations in which the same participant communicated with multiple strangers and friends.

2.1 Participants and Pairs of Participants

Participants were recruited in two ways. First, eight friend groups were recruited. Each group consisted of four participants who were friends in a school and a workplace. All participants in the four groups were male, and all participants in the other four groups were female. Therefore, the total number of participants based on this recruitment method was 32 (16 males). Second, 37 participants (19 males) who were not acquainted with participants in the friends group were recruited. All participants were Japanese speakers.

The reason for recruiting participants who were not acquainted with participants in friend groups was to collect both first-meeting (FM) and friend (FR) conversations from the same participants. First, each participant in a friend group was paired with other participants in their group. Since each participant had three friends, a total of 48 pairs of participants were obtained for FR. Then, 32 participants in friend groups were paired with three participants who were not acquainted with them. Therefore, a total of 96 pairs of participants were obtained for FM. Table 1 summarizes the statistics of this dataset.

2.2 Conversation Setting

A pair of participants communicated with each other in different rooms through a video communication system. Three conversations were recorded based on different conversation topics. Therefore, the FM dataset consisted of 288 conversations; the FR dataset consisted of 144 conversations. Each conversation lasted 20 min, and participants reported rapport ratings for their conversation partner

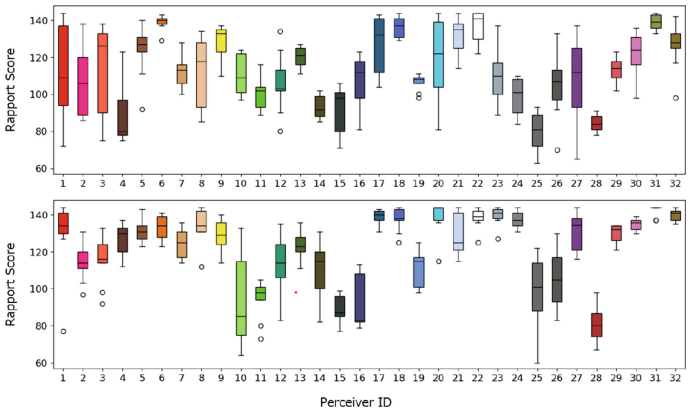

Fig. 1. A boxplot of subjective rapport ratings for each perceiver (**Upper**: first-meeting conversation, **Lower**: friend conversation).

after every conversation. The questionnaire used in our study was proposed by Bernieri et al. [14] to measure the degree of subjective rapport. They rated each item on an 8-point Likert scale. We summed the values of the 18 items after the values of the negative questions were reversed. We defined a *rapport score* as the total score.

To collect conversations with different degrees of subjective rapport from the same pair of participants, three conversations were recorded for each pair of participants based on different conversation topics: 1) self-introduction, 2) emotional episode, and 3) disclosing self-shortcomings. We selected three topics to help pairs of participants develop rapport through self-disclosure.

2.3 Conversation Topic and Rapport Score

To examine the statistical significance in the mean rapport scores between the three topics, we conducted post hoc comparisons using a t test with Bonferroni correction (the significance level was $p < 0.001$). We calculated the mean rapport score of the 32 perceivers common to both FM and FR.

In FM and FR, the mean rapport scores increased as the number of conversations increased. In FM, the mean values of the first (self-introduction), second (emotional episode), and third (self-shortcomings) topics were 107.08 (SD = 20.97), 113.3 (SD = 19.67), and 118.66 (SD = 18.84), respectively. The mean value of the first topic was significantly different than that of the second topic ($t = 6.75$, $p = 0.00$, $df = 95$); the second topic was also significantly different than the third topic ($t = 4.79$, $p = 0.00$, $df = 95$). In FR, the mean values of the first, second, and third topics were 118.01 (SD = 21.22), 121.61 (SD = 19.53), and 124.26 (SD = 20.72), respectively. However, a significant difference could not be found between the mean values of the first and second topics ($t = 2.11$, $p = 0.03$, $df = 95$) and between the second and third topics ($t = 1.40$, $p = 0.17$, $df = 95$).

There are two reasons for the increasing rapport in FM. First, participant comfort with their conversation partners increased as their total conversation times increased due to the exposure effect [15]. Second, rapport between participants increased because the conversation topics required more self-disclosure as the number of conversations increased. A previous study showed that self-disclosure contributes to rapport building [16].

2.4 Variability of Rapport Scores for Each Perceiver

The upper and lower boxplots in Fig. 1 demonstrate the variability of the rapport score for each perceiver on FM and FR, respectively. The figure shows individual differences in the tendency to rate subjective rapport. In both FM and FR, some perceivers rated rapport locally, while others rated it broadly. In addition, median values varied across participants. Compared to FM, perceivers rated rapport more highly and locally in FR. The mean values of mean rapport scores for each perceiver were 113.02 (SD = 11.87) and 121.46 (SD = 9.34) in FM and FR, respectively. The mean standard deviation values for each perceiver were 11.87 (SD = 5.70) and 9.34 (SD = 5.07).

3 Method

3.1 Problem Definition

The problem addressed in our study is to rank conversations according to the rapport score using perceivers' nonverbal features for each perceiver. Here, $\mathcal{C}_i = [c_{ijk} \mid j \in T_i, k \in [1,2,3]]$ is defined as a list containing all conversations of perceiver i, where T_i is the set containing all targets of perceiver i, and k expresses the k-th conversation topic. Each list $\mathcal{C}_i$ is associated with a list of perceiver's features $\mathcal{X}_i = [\boldsymbol{x}_{ijk} \mid j \in T_i, k \in [1,2,3]]$ and a list of rapport scores $\mathcal{Y}_i = [y_{ijk} \mid j \in T_i, k \in [1,2,3]]$. Moreover, $\boldsymbol{x}_{ijk}$ provides nonverbal features of perceiver i during their k-th conversation with target j, and y_{ijk} provides the rapport score that perceiver i gives to target j in the k-th conversation. Our goal is ranking element c_{ijk} in the conversation list C_i according to the rapport score y_{ijk}, using the perceiver's features $\boldsymbol{x}_{ijk}$ as input. For conciseness of notation, we omit ijk in c_{ijk} in the following paragraph.

To apply PL to this problem, we developed a model f that maps the perceiver's nonverbal features $\boldsymbol{x}$ to the real value $f(\boldsymbol{x})$. In the training stage, two samples were selected from each perceiver's conversation list (e.g., c_A and c_B). The model was then trained to match ordinal relations between ground-truth rapport scores (e.g., $y_A \succ y_B$) with ordinal relations between the model's output (e.g., $f(\boldsymbol{x_A}) \succ f(\boldsymbol{x_B})$). In the test stage, we obtained a predicted ranking list by ranking conversations according to the model's output.

3.2 Loss Function

We used a loss function inspired by Burges et al. [17]. Given two samples c_A and c_B, the predictive probability that c_A is higher order than c_B is given by P_{AB}:

$$P_{AB} = \frac{\exp(o_{AB})}{1 + \exp(o_{AB})}, \tag{1}$$

where $o_{AB} = f(x_A) - f(x_B)$. The true probability $\bar{P}_{AB}$ is set according to the ordinal relations between paired samples. $\bar{P}_{AB} = 1$ indicates that c_A is higher order than c_B, and vice versa. We used the cross-entropy loss function with a penalty according to the rank differences between paired samples:

$$\mathcal{L}_{AB} = \frac{\sqrt{|r_A - r_B|}}{M - 1}[-\bar{P}_{AB} \log P_{AB} - (1 - \bar{P}_{AB}) \log(1 - P_{AB})], \tag{2}$$

where r_A and r_B are the ranks of c_A and c_B in the list of conversations. M is the length of the list of conversations to which c_A and c_B belong. The loss for a paired sample with a large rank difference is higher than that for a paired sample with a small rank difference. The reason for adding a penalty was that the model emphasizes reliable paired samples. Subjective affective ratings are known to be ambiguous [11]. Paired samples with large rank differences can be considered reliable because their ordinal relationships are less likely to be reversed by variations in ratings within individuals.

3.3 Model Architecture

We developed a mapping function f inspired by Poria et al. [18]. Our mapping function was composed of unidirectional long short-term memory networks (sc-LSTM) and fully connected neural networks (FCNN).

Unimodal Mapping Function. The unimodal feature vector is given by $\boldsymbol{x}$.

$$\boldsymbol{x} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_T], \tag{3}$$

where u_t is the nonverbal features extracted during the perceiver's t-th utterances, and T is the number of perceiver utterances in a conversation. Unimodal features $\boldsymbol{x}$ are input into LSTM, and the output vector corresponding to the last utterance $\boldsymbol{g}_T$ is extracted. We then map output vector $\boldsymbol{g}_T$ to real value $f_{\mathrm{FCNN}}(\boldsymbol{g}_T)$,

$$\boldsymbol{g}_T = \mathrm{LSTM}(\boldsymbol{x}), \tag{4}$$

$$f_{\mathrm{FCNN}}(\boldsymbol{g}_T) = \mathrm{FCNN}(\boldsymbol{g}_T). \tag{5}$$

Multimodal Mapping Function. We use hierarchical multimodal fusion [18]. Consider the acoustic feature vector $\boldsymbol{x}^A$ and facial feature vector $\boldsymbol{x}^F$. Each $\boldsymbol{x}^A$ and $\boldsymbol{x}^F$ are input into different unimodal LSTM networks, and output vectors $\boldsymbol{g}^A$ and $\boldsymbol{g}^F$ are connected for each utterance.

$$\boldsymbol{g}^A = \mathrm{LSTM}^A(\boldsymbol{x}^A), \tag{6}$$

$$\boldsymbol{g}^F = \mathrm{LSTM}^F(\boldsymbol{x}^F), \tag{7}$$

$$\boldsymbol{g}^{AF} = \boldsymbol{g}^A \oplus \boldsymbol{g}^F = [\boldsymbol{g}_1^A \oplus \boldsymbol{g}_1^F, \boldsymbol{g}_2^A \oplus \boldsymbol{g}_2^F, \cdots, \boldsymbol{g}_T^A \oplus \boldsymbol{g}_T^F]. \tag{8}$$

Multimodal vector $\boldsymbol{g}^{AF}$ is input into multimodal LSTM, and the output vector corresponding to the last utterance $\boldsymbol{h}_T$ is extracted. We map the output vector $\boldsymbol{h}_T$ to the real value $f_{\mathrm{FCNN}}(\boldsymbol{h}_T)$.

$$\boldsymbol{h}_T = \mathrm{LSTM}^{AF}(\boldsymbol{g}^{AF}), \tag{9}$$

$$f_{\mathrm{FCNN}}(\boldsymbol{h}_T) = \mathrm{FCNN}(\boldsymbol{h}_T). \tag{10}$$

3.4 Feature Extraction

We extracted acoustic and facial features from each conversation. We did not use linguistic features because the conversation topic was associated with the rapport score (see Sect. 2.3). If the model has access to the content of the conversation, the model may estimate the conversation topic instead of the rapport.

Acoustic Features. We used OpenSMILE software [19] to extract acoustic features from each utterance. The acoustic features corresponded to eGeMAPS [20]. Acoustic features consisted of 88 features and were standardized for each person.

Facial Features. We used OpenFace software [21] to extract the intensity of 17 action units (AUs) from each frame. Facial features were created by computing 14 statistics from the frames corresponding to each utterance. These 14 statistics are as follows: the mean, median, standard deviation, skewness, kurtosis, maximum and minimum values, mean of the first and second differences, range, slope, intercept of the linear approximation, and 25th and 75th percentile values. Facial features thus consisted of 238 features and were standardized for each person.

4 Experiment

4.1 Comparison Model

We developed regression models to compare the PL model. For regression, we standardized rapport scores (objective variable) in two different ways. First, all rapport scores were standardized (All-perceivers), so no individual differences were addressed in the subjective ratings. Second, the rapport scores were standardized for each perceiver (Single-perceiver), which may alleviate individual differences in subjective ratings when the distribution of rapport scores shifts among perceivers. The model architecture of the regression model was the same as that of the PL model. However, pointwise learning and the root mean square error (RMSE) loss were used for regression.

4.2 Experimental Procedure and Evaluation Metrics

We defined the main-perceiver participants as the 32 participants who participated in both first-meeting (FM) and friend (FR) conversations. The training and test sets consisted of the conversations of these main-perceivers. In FM conversations, the rapport ratings of 37 participants other than the main-perceiver participants were also reported. These participants were defined as sub-perceiver participants, and their conversations were included in the training sets for FM.

We evaluated the model by double cross-validation. We considered four main-perceiver participants who were friends in one group. For outer cross-validation, we applied leave-two-groups-out cross-validation. Outer cross-validation was used to evaluate the generalization performance of the model. Next, we applied inner cross-validation to train sets obtained from outer cross-validation. For inner cross-validation, we applied leave-one-group-out cross-validation. The inner cross-validation result determined the hyperparameters used in outer cross-validation. Two cross-validation tasks ensured that the same participant's conversation was not duplicated across the training, validation, and test sets. In our study, all experiments were conducted three times based on different seed values, and their average performance was reported as the experimental results.

For the learning models, the drop rate was set to 0.25, the batch size was set to 32, and the number of epochs was set to 40. The learning rate was determined by hyperparameter optimization. Three learning rates were explored: [$5e^{-6}$, $1e^{-5}$, $5e^{-5}$]. The number of units (unimodal/multimodal LSTM hidden- and output-layer, FCNN hidden-layer) is 128.

To evaluate the ranking performance of the models, we calculated Kendall's tau correlation coefficient (KTCC) and precision at the top 3/bottom 3 (P@3/P@-3). KTCC measures the correlation between the predicted ranking list and the ground-truth ranking list. P@3/P@-3 measures how many ground-truth top-3/bottom-3 samples are present in the predicted top-3/bottom-3 samples of a model.

5 Result

For the first-meeting (FM) and friend (FR) conversations, we show the ranking performance of the models separately. First, we compare two regression models in which the rapport scores were standardized in different ways. Second, we investigate whether PL improves ranking performance and whether it is superior to regression. Table 2 indicates the ranking performance of regression and PL. The random baseline is the average ranking performance calculated 10k times between the random and ground-truth ranking lists. Bold values represent the best performances among each modality. The asterisk denotes the best performances across modalities.

Table 2. Experimental Results

Modal	Model	Standardization	First-meeting Conv.			Friend Conv.		
			KTCC	P@3	P@-3	KTCC	P@3	P@-3
A	PL	—	**0.14***	**44.44***	**37.99**	0.00	30.21	32.99
	Regression	All-perceiver	0.06	39.43	36.92	0.00	31.60	**35.07**
		Single-perceiver	0.09	43.37	35.13	**0.05**	**35.42**	34.72
F	PL	—	**0.06**	**39.43**	**44.84***	**0.10**	**39.24**	**42.36**
	Regression	All-perceiver	0.06	36.20	44.44	0.03	36.11	36.11
		Single-perceiver	0.05	32.98	43.01	-0.04	32.29	31.94
A+F	PL	—	0.06	37.28	39.07	**0.12***	**42.36***	**42.71***
	Regression	All-perceiver	0.05	34.77	**41.58**	0.06	37.85	37.85
		Single-perceiver	**0.08**	**37.99**	**41.58**	0.07	41.68	39.24
Random			0.00	33.97	33.97	0.00	34.20	34.20

Bold values represent the best performances among each modality. The asterisk denotes the best performances across modalities.

5.1 First-Meeting Conversations

In this section, we focus on the FM experimental results. As Table 2 shows, we did not observe consistent results that the ranking performance of regression (single-perceiver) was greater than that of regression (all-perceiver) regardless of modality.

The results show that PL is more effective than regression. PL achieved the best performance across modalities for all evaluation metrics (see asterisk). In KTCC and P@3, PL (A) outperformed the regression; in P@-3, PL (F) outperformed the regression. Furthermore, for both unimodal features, the performance of PL was greater than that of regression for all evaluation metrics (see bold). Regarding multimodal features, however, the performance of PL was lower than that of regression for all evaluation metrics. Next, we can see that the model using acoustic features achieved substantially higher performance in retrieving higher-ranking conversations than lower-ranking conversations; for example, PL(A) achieved higher performance for P@3 than P@-3 (P@3 = **44.44**%/P@-3 = 37.99%). In contrast, models using facial features achieved substantially higher performance in retrieving lower-ranking conversations; for example, PL(F) achieved higher performance for P@-3 than P@3 (P@3 = 39.43%/P@-3 = **44.84**%).

5.2 Friend Conversations

In this section, we focus on the experimental results in FR. As Table 2 shows, whether regression (Single-perceiver) achieved higher ranking performance than regression (All-perceiver) depended on the modality.

The results show that PL is more effective than regression. PL (A+F) achieved the best performance across modalities for all evaluation metrics (see asterisk). Furthermore, for facial features, the performance of PL was greater

than that of regression for all evaluation metrics (see bold). However, for acoustic features, the performance of PL was lower than that of regression for all evaluation metrics.

6 Discussion

6.1 Comparison of the Two Standardization Methods

In both conversations, we did not observe consistent results that the performance of regression (Single-perceiver) was greater than that of regression (All-perceiver) regardless of modality. Assuming that the distribution of rapport score shifts among perceivers, standardization (Single-perceiver) alleviates individual differences in subjective ratings. The results, therefore, imply that the distribution of rapport scores differs among perceivers. For example, the rapport score by a perceiver with a midpoint response style has an unimodal distribution, while rapport scores by a perceiver with an extreme response style have a bimodal distribution.

6.2 Comparison of Pairwise Learning and Regression

In both conversations, the PL model achieved the best performance across modalities for all evaluation metrics. The most likely explanation is that PL prevents individual differences and ambiguity in subjective ratings.

In FM conversations, PL (A) achieved higher performance than PL (F); in FR conversations, PL (F) achieved higher performance than PL (A). The result implies that differences in acoustic features between paired samples are clearer according to rapport than those in facial features in FM, and differences in facial features between paired samples are clearer according to rapport than those in acoustic features in FR.

6.3 Retrieving Higher/Lower-Ranking Rapport Conversations

In FM, models using acoustic features achieved substantially higher performance in retrieving higher-ranking conversations than lower-ranking conversations. In contrast, models using facial features achieved substantially higher performance in retrieving lower-ranking conversations. The result suggests that high rapport is encoded in acoustic features rather than facial features; low rapport is encoded in facial features rather than acoustic features. Regarding low rapport, the result is in line with a previous study finding that facial features are more indicative of low rapport than other nonverbal features (e.g., acoustic features) [3].

6.4 Limitation and Future Work

We demonstrated effective modalities for predicting subjective rapport in both FM and FR. However, we did not reveal what behavior patterns within each

modality are effective for FM and FR. In future work, the next step will be to investigate how effective behavior patterns for predicting subjective rapport differ between FM and FR.

Furthermore, we can also improve models so that models can account for interspeaker influences on nonverbal behavior. We developed a model for predicting the subjective rapport based on the nonverbal behavior of one of the speakers (perceiver) in a conversation. This method has practical advantages because the system cannot always have access to the nonverbal behaviors of the conversation partner in an online conversation. However, because interspeaker influences of nonverbal behavior (e.g., synchrony) are an important cue indicating rapport [1], models may achieve higher performance when models can have access to such cues.

7 Conclusion

We addressed predicting subjective rapport using pairwise learning (PL) in both first-meeting (FM) and friend (FR) conversations. First, we collected a dataset composed of online dyadic conversations containing subjective rapport ratings. In our dataset, the same participant communicated with multiple strangers and friends. Analysis of rapport ratings provides evidence to support that subjective rapport ratings have individual differences. Second, we investigated whether PL improves predictive performances of subjective rapport and whether it is superior to regression in not only FM conversations but also FR conversations. Experimental results demonstrated that PL is a more appropriate approach than regression for predicting subjective rapport in both FM and FR conversations. Finally, we reported effective modalities for predicting subjective rapport using PL. In FM conversations, PL models using acoustic features achieved the best performance for Kendall's tau correlation coefficient (KTCC); In FR conversations, PL models using multimodal (acoustic and facial) features achieved the best performance for KTCC. Furthermore, experimental results indicated that acoustic features are effective for retrieving high rapport conversations in FM conversations. In contrast, facial features are effective for retrieving low rapport conversations in FM conversations.

References

1. Tickle-Degnen, L., Rosenthal, R.: The nature of rapport and its nonverbal correlates. Psychol. Inq. **1**(4), 285–293 (1990)
2. Hagad, J.L., Legaspi, R., Numao, M., Suarez, M.: Predicting levels of rapport in dyadic interactions through automatic detection of posture and posture congruence. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing (2011)
3. Müller, P., Huang, M.X., Bulling, A.: Detecting low rapport during natural interactions in small groups from Non-Verbal behaviour. In: 23rd International Conference on Intelligent User Interfaces (2018)

4. Zhao, R., Sinha, T., Black, A.W., Cassell, J.: Socially-aware virtual agents: automatically assessing dyadic rapport from temporal patterns of behavior. In: Traum, D., Swartout, W., Khooshabeh, P., Kopp, S., Scherer, S., Leuski, A. (eds.) IVA 2016. LNCS, vol. 10011, pp. 218–233. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47665-0_20
5. Madaio, M., Lasko, R., Ogan, A., Cassell, J.: Using temporal association rule mining to predict dyadic rapport in peer tutoring. In: Educational Data Mining (2017)
6. Cerekovic, A., Aran, O., Gatica-Perez, D.: Rapport with virtual agents: what do human social cues and personality explain? IEEE Trans. Affect. Comput. **8**(3), 382–395 (2017)
7. Hayashi, T., et al.: A ranking model for evaluation of conversation partners based on rapport levels. IEEE Access **11**, 73024–73035 (2023)
8. Kenny, D.A.: Interpersonal Perception: The Foundation of Social Relationships. Guilford Publications (2019)
9. Baumgartner, H., Steenkamp, J.-B.: Response styles in marketing research: a cross-national investigation. J. Mark. Res. **38**, 143–156 (2001)
10. Kumano, S., Nomura, K.: Multitask item response models for response bias removal from affective ratings. In: 8th International Conference on Affective Computing and Intelligent Interaction (ACII) (2019)
11. Metallinou, A., Narayanan, S.: Annotation and processing of continuous emotional attributes: challenges and opportunities. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) (2013)
12. Lotfian, R., Busso, C.: Practical considerations on the use of preference learning for ranking emotional speech. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5205–5209 (2016)
13. Parthasarathy, S., Lotfian, R., Busso, C.: Ranking emotional attributes with deep neural networks. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2017)
14. Bernieri, F.J., Gillis, J.S., Davis, J.M., Grahe, J.E.: Dyad rapport and the accuracy of its judgment across situations: a lens model analysis. J. Pers. Soc. Psychol. **71**(1), 110–129 (1996)
15. Zajonc, R.B.: Attitudinal effects of mere exposure. J. Pers. Soc. Psychol. **9**(2p2), 1–27 (1968)
16. Zink, K.L., et al.: "Let me tell you about my…" provider self-disclosure in the emergency department builds patient rapport. West. J. Emerg. Med. **18**(1), 43–49 (2017)
17. Burges, C., et al.: Learning to rank using gradient descent. In: Proceedings of the 22nd International Conference on Machine Learning (2005)
18. Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., Morency, L.-P.: Context-dependent sentiment analysis in user-generated videos. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (2017)
19. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the Munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM International Conference on Multimedia (2010)
20. Eyben, F., et al.: The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comput. **7**(2), 190–202 (2016)
21. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.-P.: Openface 2.0: facial behavior analysis toolkit. In: 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018) (2018)