Multimodal Technologies and Interaction

MDPI

Article

Influence of Personality Traits and Demographics on Rapport Recognition Using Adversarial Learning

Wenqing Wei [1], Sixia Li [1], Candy Olivia Mawalim [1], Xiguang Li [1], Kazunori Komatani [2] and Shogo Okada [1,*]

[1] Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Nomi 923-1292, Japan; s2020011@jaist.ac.jp (W.W.); lisixia@jaist.ac.jp (S.L.); candylim@jaist.ac.jp (C.O.M.); s2020425@jaist.ac.jp (X.L.)

[2] Department of Knowledge Science, Institute of Scientific and Industrial Research, Osaka University, Ibaraki 567-0047, Japan; komatani@sanken.osaka-u.ac.jp

* Correspondence: okada-s@jaist.ac.jp; Tel.: +81-076-151-1205

Abstract: The automatic recognition of user rapport at the dialogue level for multimodal dialogue systems (MDSs) is a critical component of effective dialogue system management. Both the dialogue systems and their evaluations need to be based on user expressions. Numerous studies have demonstrated that user personalities and demographic data such as age and gender significantly affect user expression. Neglecting users' personalities and demographic data will result in less accurate user expression and rapport recognition. To the best of our knowledge, no existing studies have considered the effects of users' personalities and demographic data on the automatic recognition of user rapport in MDSs. To analyze the influence of users' personalities and demographic data on dialogue level user rapport recognition, we first used a Hazummi dataset which is an online dataset containing users' personal information (personality, age, and gender information). Based on this dataset, we analyzed the relationship between user rapport in dialogue systems and users' traits, finding that gender and age significantly influence the recognition of user rapport. These factors could potentially introduce biases into the model. To mitigate the impact of users' traits, we introduced an adversarial-based model. Experimental results showed a significant improvement in user rapport recognition compared to models that do not account for users' traits. To validate our multimodal modeling approach, we compared it to human perception and instruction-based Large Language Models (LLMs). The results showed that our model outperforms that of human and instruction-based LLM models.

Keywords: multimodal dialogue systems; user rapport; adversarial learning; users' personal traits

Academic Editor: Mark Billinghurst

Received: 13 November 2024
Revised: 16 January 2025
Accepted: 4 February 2025
Published: 20 February 2025

Citation: Wei, W.; Li, S.; Mawalim, C.O.; Li, X.; Komatani, K.; Okada, S. Influence of Personality Traits and Demographics on Rapport Recognition Using Adversarial Learning. *Multimodal Technol. Interact.* **2025**, *9*, 18. https://doi.org/10.3390/mti9030018

1. Introduction

With recent advances in natural language processing and speech recognition technologies, spoken dialogue systems such as Amazon Alexa, Siri, and Google Assistant have become widely popular across various domains. In recent years, dialogue systems LLMs-based, such as ChatGPT, Gemini, and Claude, have become mainstream in research and applications. Their advanced language generation and natural interaction capabilities have further driven the progress of research and development in non-task oriented dialogue systems [1–4].

Performance evaluation of dialogue systems plays a crucial role in optimizing data-driven dialogue systems and has been an active area of research. Rapport, a widely used evaluation aspect of dialogue systems, has been defined as mutual attention, positivity,

and coordination [5]. It refers to a harmonious, understanding, and trusting relationship established between individuals or between humans and machines. While rapport in human interactions [6–8] has been extensively studied, research on building and maintaining rapport in human-computer interactions [9] is relatively recent. Establishing good rapport in human-computer interaction can make these interactions more natural and effective, thereby increasing user satisfaction and experience. By evaluating user rapport during or after interacting with a dialogue system, developers can gain insights into how users respond to the system.

Given this background, in this study, we first investigate the impact of users' personal traits, such as age, gender, and personality, on user rapport recognition. We employed a dataset containing 18 types of user rapport with personal information such as age, gender, and personality traits, as shown in Figure 1. This dataset incorporated multiple modalities, including audio, body motion, visual cues, and transcript data, providing a comprehensive basis for evaluating the user rapport of dialogue systems. Following related research [10,11], we used the Big Five as the source of user personalities in this dataset. In this way, this dataset allowed us to explore the impacts of user personal information on user rapport recognition. Through the analysis in Section 3.3, we confirmed that users' personal traits significantly affect rapport recognition.

While these traits may offer some degree of relevance, an overreliance on such features can lead to potential biases in the model, which may affect its fairness and performance across different user groups. Therefore, to mitigate the influence of personal information on the model, we employ an adversarial learning method. This approach uses gradient reversal techniques, which aim to reduce the model's dependence on user traits, allowing it to focus on more general and crucial features. As a result, this method enhances the robustness and accuracy of user rapport recognition, ensuring that the model remains adaptable to various user profiles and scenarios. To validate the effectiveness of the proposed ADVER-based model Sections 7.2 and 7.3 compare the results of the adversarial approach, the baseline, human model, and instruction-based LLM methods. The results demonstrated that the adversarial approach achieved superior performance. The main contributions of this study can be summarized as follows:

We first addressed the research question (RQ1): "Does the adversarial learning of users' personal information contribute to rapport recognition performance?" The effectiveness of the specifically utilized adversarial learning method is discussed in Sections 7.1 and 8.2.

Age, gender, and personality influence users in human-computer dialogue in different ways. Age and gender often result in significant expression differences but typically do not directly impact user interaction with the employed system. Personality, however, can influence how users interact. For example, extroverted individuals may prefer open-ended conversations and more frequent interactions, whereas introverted individuals may favor direct and concise communication. Our second research question (RQ2) was as follows: "Are there differences between the impacts of demographic data (age and gender) and personality on user rapport recognition?" The relevant research findings are discussed in Section 7.2.

Additionally, the small size of our target dataset, with a total of 125 dialogues, might affect the performance of the developed model. To validate the effectiveness of the baseline and proposed models, we compared the recognition results of the machine learning models with user rapport scores annotated by third-party experts and instruction-based LLM. Our third research question (RQ3) was as follows: "Do machine learning models outperform the results derived from multiple human observations and instruction-based LLM in terms of estimation accuracy?" The comparison results are detailed in Section 7.3.

SYS: Since last year, due to the impact of self–restraint on going out, the opportunity for dining out has decreased. Have you been creative in your meals?
(去年から外出自粛の影響で、外食があまりできないですが、食事で何か工夫されていることはありますか？)
USER: I've tried to add some variations to the dishes at home.
家で作る料理のバリエーションを増やしました
SYS: Have you tried adapting recipes or something similar?
(アレンジレシピとかはありますか？)
USER: Basically, I find recipes online and make various dishes, so I follow the recipe instructions without making too many adjustments.
(基本的にはネットで探して |変わった料理を作っているので アレンジもレシピ通りに作ってて自分では何もしていません)
SYS: I see. I tried a recipe using canned mackerel, and I was able to make a delicious Aqua Pazza. You can find it if you search online.
(なるほど、私はサバ缶を使ったレシピで、アクアパッツァがなかなかおいしくできました。ググったら出てきます。)
USER: Oh, I see. It looks very delicious.
(あーそっかとてもおいしそうですね)
SYS: It was easy to do while watching the video.
(動画をみながら作ると、簡単でした)
USER: Mm-hmm, it seems simple and like a good recipe.
(簡単そうでいいレシピですね)

User traits :{
"age" : " 20s",
" gender" : "female",
"Big-five": "4, 3, 3, 4, 5",
}

Figure 1. An example dialogue session contained in our dataset(Contains the English translation along with the corresponding original data in Japanese).

2. Related Work

Generally, dialogue systems can be classified into two categories: tasks-oriented [12–14] and non-task-oriented [15,16] systems. Task-oriented dialogue systems are designed to help humans achieve desired goals. These systems focus on tasks such as booking flights, ordering food, or making hotel reservations. Their goal is to offer the best possible solutions by asking relevant questions and providing informative responses. Non-task-oriented dialogue systems are designed for general conversations and do not have any specific goals or objectives. These systems aim to engage users in casual conversation and provide them with enjoyable experiences.

Given the diversity and complexity of dialogue systems, evaluating intelligent assistants has become a challenging task and an active research field. Most statistical approaches to spoken dialogue assessment consider objective criteria such as dialogue length or task success rates [17]. However, these metrics do not necessarily correspond to the subjective and immediate user rapport of the target conversation. Especially for non-task-oriented conversations such as small talk and multidomain dialogues, no task success information is available when interacting with simulated or recruited users. This lack of information makes it difficult to evaluate non-task-oriented dialogue systems.

To address this problem, researchers have recently focused on more user-centered criteria, such as measuring user rapport levels during or after interacting with a dialogue system. Engelbrecht et al. and Klaus-Peter et al. [18] used hidden Markov models (HMMs) to recognize user satisfaction at each step of a dialog. With the continuous advancement of neural networks, numerous researchers are using various deep-learning techniques to predict user satisfaction. For example, Ultes et al. [19] proposed a bidirectional long short-term memory (BiLSTM) model to assess the quality of interactions and achieve improved performance. To capture the different aspects of user satisfaction, ref. [20] proposed a multitask deep learning-based neural network model that predicts user sentiment, user interest, and user topic continue based on the exchange level. T.E.Kim et al. [21] proposed a model that combines the user-utterance generation task with the user satisfaction scoring and action prediction tasks by applying a deep multitask neural model to achieve good user satisfaction prediction performance. A good dialogue system should provide coherent and appropriate responses and be sufficiently engaging to leave an overall rapport with the user. Therefore, it is essential to analyze the user rapport of the dialogue system not only at the exchange level but also at the dialogue level. The primary objective of the dialogue-level user satisfaction evaluation task involves learning dialogue strategies that maximize impressions in an overall conversation, which also helps identify problematic conversation topics that lead to user dissatisfaction. Ref. [22] used a statistical classification method with support vector machines to predict interaction quality at the dialogue level with field and laboratory data, thus overcoming the limitation of using task success.

To improve the non-task-oriented multimodal dialogue system, Wei et al. [23] used automatic multimodal features to evaluate such systems at the dialogue level. Furthermore, to employ the relationship between user impressions at the dialogue and exchange levels, Bodigutla et al. [24] proposed a multitask base model that jointly predicts turn-level annotation labels and user impression level for dialogue. Given the impressive reasoning and dialogue understanding capabilities demonstrated by LLMs, researchers have also employed them to evaluate user performance at the exchange level [25,26]. Zheng et al. [27] utilize large language models (LLMs) as judges to evaluate multi-turn dialogues. More recently Md Tahmid Rahman Laskar et al. [28] provide a systematic review of the main challenges in evaluating LLMs, including issues of reproducibility, reliability, and robustness.

Importantly, most existing methods focus primarily on algorithmic improvements without considering the impact of users' personal information on user rapport. Several notable works have recently demonstrated that adversarial methods are successful in terms of enhancing the robustness and generalizability of models in various tasks. Meng et al. [29] utilized an adversarial speaker adaptation method to achieve improved speech recognition in Microsoft short message tasks by aligning the features of speaker-dependent models with a reference model, achieving significant word error rate gains. Gao et al. [30] used adversarial domain adaptation and a center loss to enhance the generalization capabilities of cross-corpus speech emotion recognition systems.

This study aims to investigate and mitigate the impacts that may lead to potential biases in the model of user rapport recognition for dialogue systems. Inspired by previous works [23,29], we examined the relationships between 18 rapport labels and users' personal information. We subsequently employed an adversarial-based model designed to more effectively adapt to these personal information variations. An overview of this study is presented in Figure 2.

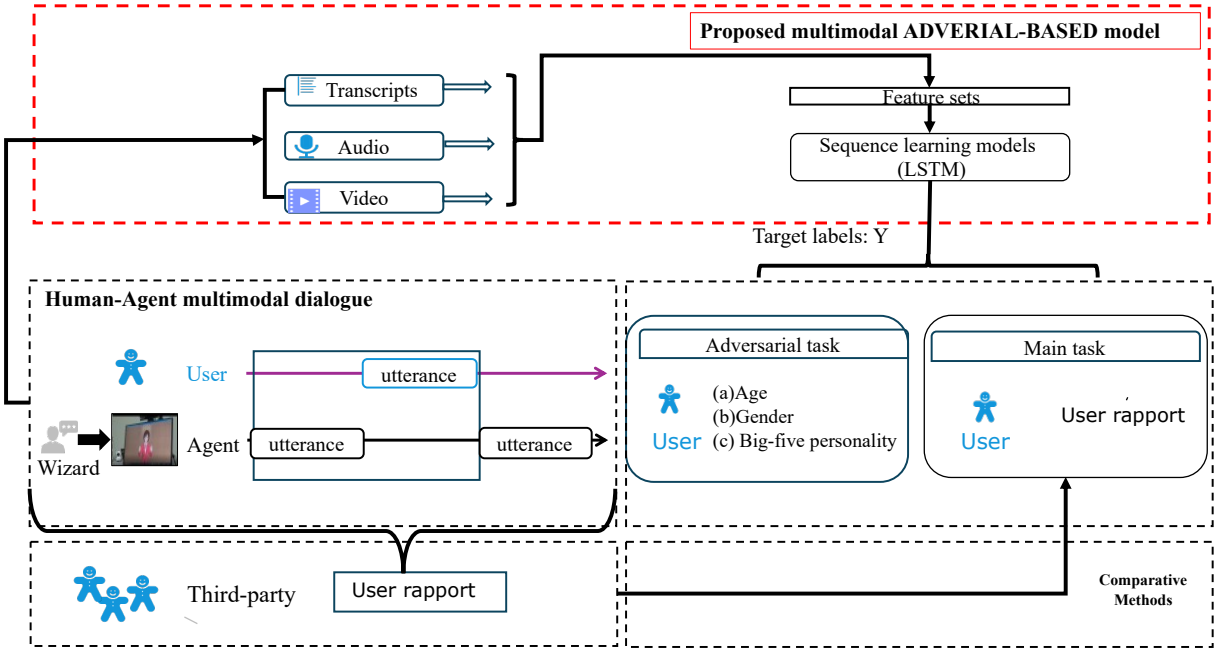

Figure 2. Overview of the multimodal model for adapting users' traits to recognize user rapport.

3. Dataset

3.1. Data

To develop a user-adaptive multimodal dialog system, Komatani et al. [31] collected the multimodal Hazumi dataset. All the Hazumi data are publicly available (https://www.nii.ac.jp/dsc/idr/rdata/Hazumi/ (accessed on 3 February 2025)), on the basis of the data policy and anyone can apply. Most previous works [23,32,33] have implemented this on the basis of dataset involving laboratory settings. To better reflect real-world conditions, this study used three multimodal dialogue datasets Hazumi2105 (https://github.com/ouktlab/Hazumi2105/ (accessed on 3 February 2025)), Hazumi2012 (https://github.com/ouktlab/Hazumi2012/ (accessed on 3 February 2025)), and Hazumi2010 (https://github.com/ouktlab/Hazumi2010/ (accessed on 3 February 2025)). Each participant was recorded online. The MMD-Agent platform [34] was used as the interface for interacting with the participants, with their responses being controlled by an operator (Wizard). All datasets were arranged to record facial videos, and audio data via microphones and cameras through the Zoom software platform (https://zoom.us/ja/download). Throughout the interactions, if participants displayed signs of disinterest, the Wizard would proactively change the topic to rekindle their engagement. Conversely, if the participants appeared interested and actively participated in the conversation, the Wizard would listen and respond. The specific details of the dataset are provided in Table 1.

Table 1. Data summary for online Hazumi datasets.

Dataset	**Hazumi Datasets** [35]		
Version Name	Hazumi2010	Hazumi2012	Hazumi2015
Overview	Dialogues between human participants and virtual agents operated by a human Wizard for approximately 15 to 20 min per dialogue		

Table 1. *Cont.*

Dataset	**Hazumi Datasets** [35]		
Instructions to the Wizard Participants	Chit-chat involving any topics to make the participants enjoy the dialogue		
Participants	33 (17 per male, 16 female)	63 (29 per male, 34 per female)	29 (14 perr male, 15 per female)
	Aged 20 to 70 (27 per: age < 50, 6 per: age > 50)	Aged 20 to 70 (54 per: age < 50, 9 per: age > 50)	Aged 20 to 70 (27 per: age < 50, 6 per: age > 50)
Sensors	Videos and voices of the participants		
	Video of agent		
Manual annotations	Third-party sentiment with respect to the exchange turn level (on a 7-point scale) provided by 5 annotators Topic continuance concerning the exchange turn level (on a 7-point scale) provided by 5 annotators		
	18 types of user rapport labels on dialogue level (on an 8-point scale), by 5 annotators and the participant		
	3 types of user rapport (coordinateness, awkwardness, and friendliness) at the dialog level (on an 8-point scale) provided by the Wizard		

3.2. Annotations

3.2.1. User Rapport

This dataset employed a questionnaire comprising 18 labels to measure the rapport of each user with the dialogue, as described in [36]. The questionnaire measured cognition and rapport in interpersonal communications. The 18 items were "well coordinated", "boring", "cooperative", "harmonious", "unsatisfying", "uncomfortably paced", "cold", "awkward", "engrossing", "unfocused", "involving", "intense", "friendly", "active", "positive", "dull", "worthwhile", and "slow". Both the dialogue users and three independent experts rated each label on an eight-point scale ranging from 1 to 8. The users' ratings were used as the ground truth. The interannotator agreement among the experts was measured using Cronbach's alpha. Additionally, the distribution of the users' dialog-level labels was analyzed, and the scores (1–8) were converted into binary categories (high and low) with a threshold of 4. The results are presented in Table 2. We found a high degree of consensus among users for some rapport labels such as "cooperative" (119/6), "friendly" (125/0), and "positive" (124/1), indicating that these system rapport labels were uniform and did not require recognition. In contrast, rapport labels that lacked consensus among the users could be beneficial for improving the dialogue system. Following previous studies [23,33], we selected rapport labels that are reflected more disagreement among users, such as "well coordinated" (102/23), "awkward" (67/58), and "engrossing" (75/50), as the prediction targets.

Table 2. Agreement scores of annotators and distribution of high/low data (with 4 as the threshold) for 18 annotation types.

User Rapport	**Hazumi2010**	**Hazumi2012**	**Hazumi2105**	**High/Low**	**Level**
well coordinated	0.804	0.774	0.631	102/23	pos
boring	0.856	0.793	0.667	23/102	neg
cooperative	0.844	0.731	0.575	119/6	pos

Table 2. *Cont.*

User Rapport	Hazumi2010	Hazumi2012	Hazumi2105	High/Low	Level
harmonious	0.823	0.711	0.643	105/20	pos
unsatisfying	0.781	0.757	0.726	17/108	neg
uncomfortably paced	0.577	0.710	0.511	72/53	neg
cold	0.716	0.546	0.386	17/108	neg
awkward	0.640	0.717	0.599	67/58	neg
engrossing	0.833	0.795	0.741	75/50	pos
unfocused	0.701	0.557	0.342	21/104	neg
involving	0.823	0.717	0.640	110/15	pos
intense	0.402	0.704	0.495	53/72	neg
friendly	0.854	0.721	0.667	125/0	pos
active	0.879	0.807	0.770	105/20	pos
positive	0.833	0.737	0.654	124/1	pos
dull	0.826	0.746	0.616	20/105	neg
worthwhile	0.794	0.716	0.612	106/19	pos
slow	0.820	0.696	0.506	50/75	neg

3.2.2. Personality

The Ten Item Personality Inventory (TIPI) is a well-validated and brief version. Research [37] has shown that this questionnaire has high reliability and validity for measuring personality. The scale can be completed in approximately 1 min and comprises the following 10 items:

I think I am:

- Q1. Energetic and outgoing
- Q2. Easily dissatisfied and prone to conflict
- Q3. Self-demanding and strict
- Q4. Anxious and worried
- Q5. Enjoys novelty
- Q6. Modest and shy
- Q7. Caring and kind
- Q8. Careless about details
- Q9. Calm and stable
- Q10. Lacks creativity, ordinary

According to [35], two items represent each of the Big Five personality traits. Manually assessed personality scores could be obtained from the participants; item scores are determined through simple calculations (with Q*i* representing the rating score for item *i*), such as subtracting the rating score for Q6 from that of Q1 to derive the extraversion score. The calculation method for each personality score can be calculated as follows:

- (1) Extraversion: (Q1 + 8 − Q6)/2,
- (2) Agreeableness:(Q2 + 8 − Q7)/2;
- (3) Conscientiousness: (Q3 + 8 − Q8)/2;
- (4) Neuroticism:(Q4 + 8 − Q9)/2;
- (5) Openness:(Q5 + 8 − Q10)/2.

3.3. Data Analysis

Owing to the influence of user age, gender, and personality on dialogues, this study aimed to investigate the impact of users' personal information on the rapport of dialogue systems. We examine the differences among user annotations based on age, gender, and personality. The dialogue-level labels depict the annotations from both positive and negative polarities. Because the labels in distinct polarities represent opposing annotations, we partitioned the dialogue-level labels into two categories to enable the presentation of different annotations. The positive category contains well-coordinated, cooperative, harmonious, engrossing, involving, friendly, active, positive, and worthwhile annotations. Conversely, the negative category includes boring, unsatisfying, uncomfortably paced, cold, awkward, unfocused, intense, dull, and slow annotations.

3.3.1. The Relationships Between User Rapport Labels and User Gender and Age

Figure 3 separately shows the average ratings of 125 participants for the 18-item post questionnaires consisting based on age (old/young) with the boundary set at 50 years old and gender (male/female) separately. The vertical axis denotes the mean user rapport rating. The horizontal axis represents the 18 questionnaire items, with the first nine items being positive labels. The overall average of the positive labels is displayed in the tenth position. Conversely, the remaining nine items are negative labels represented by tags that range from the eleventh to the nineteenth positions. The overall average of the negatively labeled items is shown at the right end of the horizontal axis. With respect to age, older participants tended to be more positive than younger participants in terms of sentiment-type rapport, such as friendly, active, positive, bored, and cold. In terms of gender, the female data were more sensitive and had higher standard deviations and means than the male data did for most labels, including positive and negative user rapport. Therefore, it is important to consider age and gender when recognizing a user's rapport at the dialogue level.

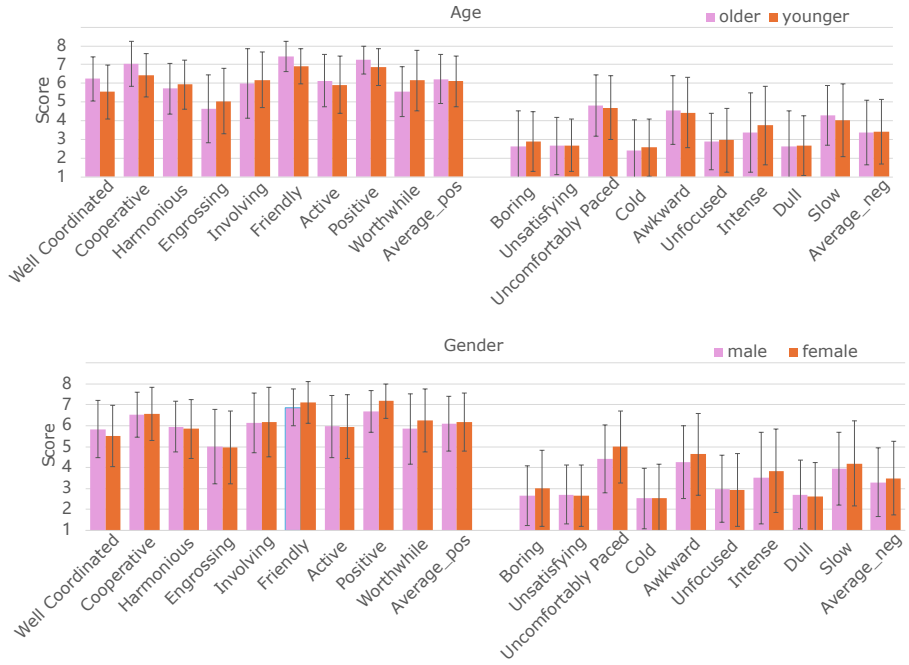

Figure 3. The average scores for 18 types of annotations for age (**top**) and gender (**bottom**). On the left side of each figure is the average score for each positive rapport label. Average_pos represents the average of all positive rapport labels. The average score for each negative rapport label is shown on the right side of each figure. Average_neg represents the average of all negative rapport labels.

3.3.2. Relationships Between User Satisfaction and User Personalities

In this section, we focus on exploring the relationships between user personality and user rapport. We calculated the mean performance of different user rapport types across user personalities. In this section, we discuss extraversion and neuroticism, as representative examples. Figure 4 illustrates the average ratings provided by 125 participants on the post questionnaires, which compriseding 18 items related to the Big Five personality traits. The personality scores were converted into two types (high personality and low personality) with a threshold of 4.

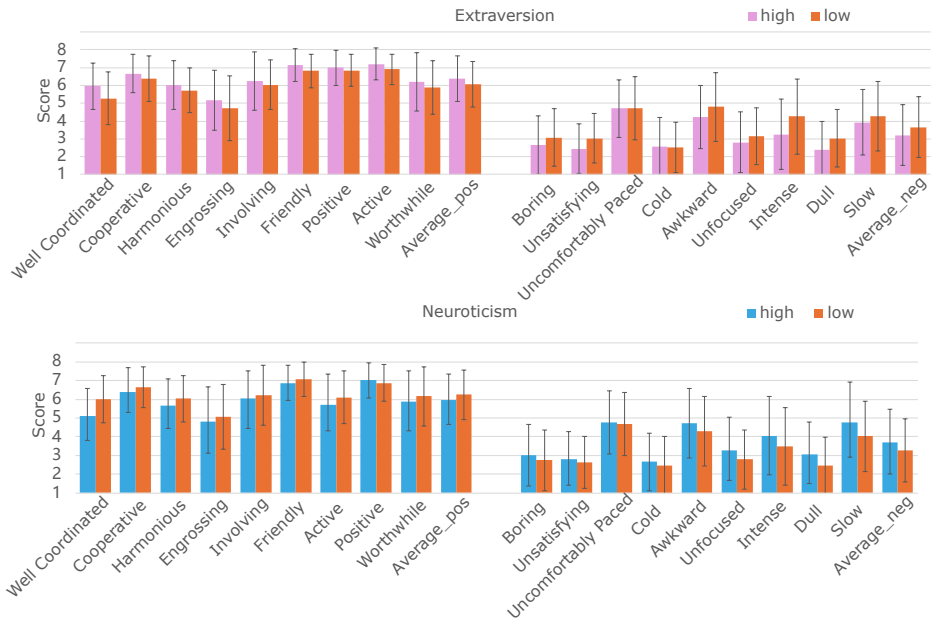

Figure 4. The average scores for 18 types of annotations for Big Five personality traits. On the left side of each figure is the average score for each positive rapport label. Average_pos represents the average of all positive rapport labels. The average score for each negative rapport label is shown on the right side of each figure. Average_neg represents the average of all negative rapport labels.

In Figure 4, blue represents the mean user rapport scores of high-personality users, whereas orange represents the mean scores of low-personality users. We observed that for positive user rapport (columns 1 to 9), in terms of extroversion, high-personality users tended to have higher mean rapport label scores than did low-personality users. Conversely, for negative user rapport labels (columns 10 to 18), high-personality users tended to have lower mean rapport scores than low-personality users did. This result indicated a positive relationship between extraversion and user rapport and a negative relationship with negative user rapport labels. Similarly, for neuroticism, we found that for positive user rapport labels, high-personality users had lower mean user rapport scores than low-personality users. Conversely, for negative user rapport labels, high-personality users tended to have higher mean rapport scores than low-personality users did. This result indicated a negative relationship between neuroticism and positive user rapport labels and a positive relationship with negative user rapport labels.

Table 3(a) lists the Pearson correlations between the Big Five personality traits and the positive dialog-level user rapport scores. Generally, a correlation coefficient above 0.1 signifies a weak correlation, whereas a correlation coefficient above 0.3 signifies a moderate correlation. Table 3(b) presents the coefficient values between the Big Five personality traits and the negative dialogue-level user rapport labels. Each row represents a personality, and each column displays a dialog-level user rapport label. The intersection of a row and column indicates the coefficient value between the personality and rapport of the user. As

illustrated in Table 3(a), all the coefficients between extraversion, conscientiousness, and openness personality traits and the positive dialog-level user rapport labels were positive. However, the coefficient between openness and the positive label was an exception, being negative but close to zero. The average coefficients between extraversion, conscientiousness, and openness and the positive dialog-level annotations were 0.205 for extraversion, 0.108 for conscientiousness, and 0.123 for openness, respectively. Conversely, the correlation coefficients between agreeableness and neuroticism personality traits and the positive user rapport labels were negative, with the average coefficients being −0.129 for agreeableness and −0.144 for neuroticism. In contrast, all the coefficients between extraversion, conscientiousness, and openness personality traits and the negative dialog-level user rapport labels were negative. This result is consistent with the results shown in Figure 4. The above analysis indicates that stronger correlations are present between user personalities and dialog-level user rapport labels.

Table 3. Pearson correlation coefficient results between user personality and dialogue-level user rapport. (a) shows the coefficients between user personality and positive dialogue-level user rapport labels. (b) shows the Pearson correlation coefficients between user personality and negative dialogue-level user rapport labels.

Label	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
(a) Pearson correlation coefficients of positive user rapport labels					
Well Coordinated	0.314	−0.164	0.118	−0.347	0.146
Cooperative	0.157	−0.112	0.099	−0.148	0.057
Harmonious	0.193	−0.084	0.158	−0.134	0.098
Engrossing	0.229	−0.031	0.089	−0.117	0.223
Involving	0.117	−0.186	0.077	−0.052	0.048
Friendly	0.228	−0.230	0.154	−0.173	0.158
Active	0.349	−0.076	0.220	−0.155	0.277
Positive	0.092	−0.260	−0.058	0.006	−0.019
Worthwhile	0.169	−0.015	0.117	−0.179	0.115
Average_pos	0.205	−0.129	0.108	−0.144	0.123
(b) Pearson correlation coefficients of negative user rapport labels					
Boring	−0.166	0.056	−0.046	0.080	−0.151
Unsatisfying	−0.241	0.133	−0.016	0.073	−0.155
Uncomfortably Paced	−0.077	−0.042	−0.118	0.076	−0.088
Cold	0.053	0.141	0.044	0.086	0.013
Awkward	−0.230	0.034	−0.098	0.171	−0.138
Unfocused	−0.178	0.015	−0.057	0.086	0.027
Intense	−0.239	0.002	0.039	0.202	−0.087
Dull	−0.237	0.110	−0.093	0.206	−0.140
Slow	−0.105	0.123	−0.281	0.192	−0.206
Average_neg	−0.158	0.063	−0.070	0.130	−0.103

4. Features Extraction

4.1. Audio Feature

For audio features, we use OpenSMILE [38] to extract exchange-level acoustic features. These features corresponded to the extended Geneva minimalistic acoustic parameter set (eGeMAPS), which excels in emotion-related fields.

4.2. Linguistic Feature

The study extracted linguistic features from the participants' utterances and dialogue log data. We extracted two types of linguistic features from the manual transcription of spoken dialogue contents:

Part of speech: The sentences were segmented into words and annotated with universal part-of-speech (POS) tags via Stanza NLP (https://github.com/stanfordnlp/stanza (accessed on 3 February 2025)).

The PoS tag set was composed of 17 types: "adjective", "adposition", "adverb", "auxiliary", "coordinating conjunction", "determine", "interjection", "noun", "numeral", "particle", "pronoun", "proper noun", "punctuation", "subordinating conjunction", "symbol", "verb", "other". The PoS categories (nouns, verbs, etc.) in a user's utterance were counted. The frequencies of the PoS categories, such as nouns and verbs, in each user's utterance were calculated. We utilized a 17-dimensional vector to represent the 17 POS tags.

BERT (bidirectional encoder representations from transformers [39]): In this study, we employed a pretrained model that specifically focused on Japanese text (trained on Wikipedia) (https://github.com/yoheikikuta/bert-japanese (accessed on 3 February 2025)). This model was utilized to extract features from the text at the exchange level. Consequently, we obtained a 768-dimensional vector representing the text representation.

4.3. Visual Feature

We extracted facial features as visual features via an RGB camera.

Facial landmark feature: OpenFace [40] software outputs the three-dimensional coordinates of 68 facial landmarks in each frame. This study chose ten facial landmarks, including 2 points on each eye, 4 points around the mouth, and 2 points on the eyebrows. We utilized the same methodology employed for tracking body features to track facial features. For each user exchange, we extracted the maximum acceleration value, as well as the maximum, mean, and standard deviation of the velocity value, resulting in facial features. Ultimately, we obtained a 40-dimensional vector representing these features.

Action units: Facial expressions play a crucial role in displaying emotional states and facilitating turn-taking during conversations. These expressions are typically represented by facial action units (AUs), which provide objective descriptions of facial muscle activations [41]. In this study, OpenFace software (https://cmusatyalab.github.io/openface/) was employed to extract 18 types of AUs, each rated as 0 (absent) or 1 (present). The average value of each AU within an exchange was then calculated to derive facial AU features (18 dimensions). Consequently, a total of 58 dimensions of facial features were utilized in this study.

5. Methods

5.1. Models

Single Task Deep Learning Neural Network (Baseline)

To capture the dynamic changes in the multimodal behaviors of users during a conversation, based on previous work [33], we used LSTM methods to evaluate dialogue-level user impressions. As described in Section 4, different unimodal features (audio a_t: 88-dim., linguistic l_t: 785-dim., and video v_t: 58-dim.) were extracted from the t-th exchange. We used the early fusion method to concatenate different unimodal features, generating the exchange-level multimodal feature $x_t = [a_t, v_t, v_t]$. The multimodal feature $X = (x_1, x_2, ..., x_t)$ was used as the input of the neural network models. For all the models with one LSTM layer and 128 units, we obtained a 128-dimensional hidden state from the recurrent layer. The recurrent layer was followed by a fully connected layer, which projected the (128-dimensional) output. At the end of the model output layer containing two units, the log-Softmax function was used to output the probabilities of different user rapport labels.

As shown in Figure 5a, the output at the final moment h_t can be regarded as a representation of the whole sequence, which uses a fully connected layer followed by a softmax nonlinear layer to predict the probability distribution over different classes. As outlined in

Section 4, various unimodal features (audio a_t: 88-dim., linguistic l_t: 785-dim., and video v_t: 58-dim.) were extracted from the t-th exchange. To combine these unimodal features, we employed the early fusion technique, concatenating them to create the exchange-level multimodal feature $x_t = [a_t, v_t, v_t]$. The multimodal feature $X = (x_1, x_2, \ldots, x_t)$ was used as the input for the neural network models.

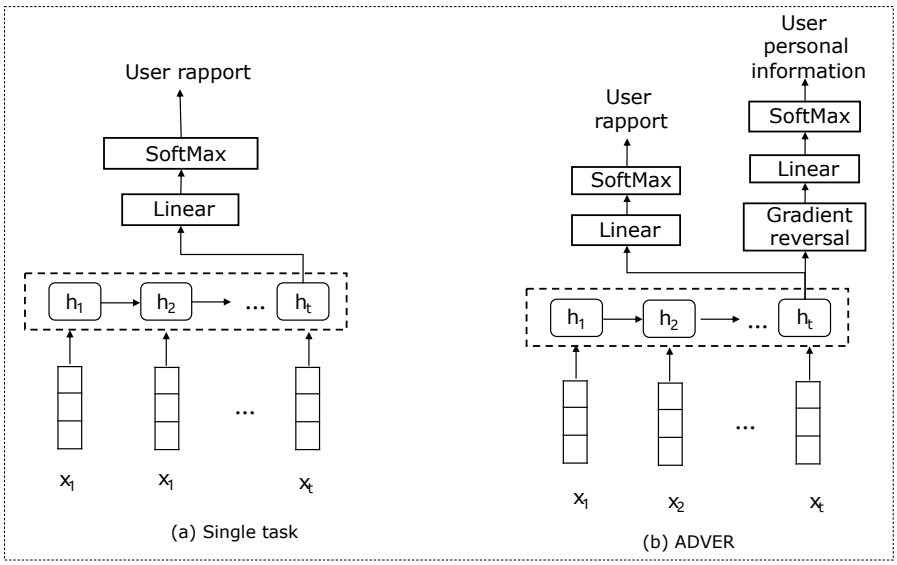

Figure 5. The structure of single task model and adversarial model.

5.2. Domain Adversarial Neural Network for User Rapport (Proposed Model)

To mitigate the influence of user traits on the recognition of user rapport, we employed an adversarial learning method [42]. This approach was selected due to its ability to reduce the model's reliance on potentially biased features, such as age, gender, and personality, ensuring fairer and more accurate rapport recognition. In this approach, a neural network model is trained to perform two tasks simultaneously: a primary task (user rapport) and a domain (user traits) adaptation task. The key idea behind gradient reversal is to force the utilized neural network to learn domain-invariant features during training. During the forward pass of the primary task, the shared layers of the network extract features from the input data. These shared layers are responsible for capturing general patterns and features across both tasks. However, during the backward pass, the gradients are reversed for the domain adaptation task. This means that the gradients flowing through the shared layers are multiplied by a negative scalar, effectively reversing their direction. As a result, the shared layers are encouraged to learn features that are domain-invariant, making them less sensitive to variations between different domains. In summary, gradient reversal allows a neural network to learn task-specific representations while simultaneously learning domain-invariant features. This approach helps improve model performance in the primary task by reducing the influence of domain-specific characteristics.

In the ADVER-based model, we used single LSTM layers with 128 units as the shared layers. These layers extracted features at both the user personal information level and the dialog level of user rapport for the tasks shown in Figure 5b. For the domain adaptation task, we obtained 128-dimensional hidden states H = (h_1, h_2, ..., h_t) from the LSTM layers, and the output at the final moment h_t could be regarded as a representation of the whole sequence. Subsequently, the h_t served as the input of a gradient reversal layer, yielding an output G, which was then used as the input of a fully connected layer followed by a Softmax nonlinear layer for predicting the probability distribution over different classes of user' personal information. For the dialog-level user rapport recognition task, the structure

was the same as that used for the single task, and the mathematical formula of the model can be described as follows:

$$\text{Share layer: } h_t = \text{LSTM}(\mathrm{x}_t W_e, \mathrm{h}_{t-1}) \tag{1}$$

$$\text{Gradient reversal layer: } G = \text{GRL}(h_t W_g + b_g) \tag{2}$$

$$\text{Adversarial task layer: } U = \text{Softmax}(G W_u + b_u) \tag{3}$$

$$\text{User rapport task layer: } D = \text{Softmax}(h_t W_d + b_d) \tag{4}$$

The loss of the ADVER base model can be defined as shown in Equation (5), where L_d is the cross-entropy loss for the dialogue user rapport classifier, and L_u is the cross-entropy loss for the users' personal information classifier. These two classifiers were adversarially trained. Specifically, the model parameters for user rapport classification were adjusted to minimize L_d, and the users' personal information classification was adjusted to maximize L_u. Minimax competition enhances the discriminability of user rapport, and suppressed the discriminability of users' personal information, leading the model to converge to a state where the embeddings we extracted could recognize user rapport but could not correctly classify users' personal information. Therefore, under ideal conditions, the embeddings we obtained were not influenced by users' personal information.

$$L = L_d - \lambda * L_u \tag{5}$$

6. Experiments

User rapport recognition is a time series task that requires time series information to achieve improved model performance. Therefore, this study used recurrent neural networks that are suitable for handling sequential inputs consisting of time series information. Based on previous research findings, we used long short-term memory (LSTM) as the baseline model to capture sequences of multimodal behaviors. Additionally, we employed adversarial models to eliminate the impact of users' personal information. In particular, as previously indicated, we aimed to answer the following research questions.

- **(RQ1):** Does the adversarial learning of users' personal information contribute to rapport recognition performance?
- **(RQ2):** Are there differences between the impacts of demographic data (age and gender) and personality on user rapport recognition?
- **(RQ3):** Do machine learning models outperform the results derived from multiple human observations and instruction-based LLM in terms of estimation accuracy?

6.1. Experimental Settings

Given the relatively small dataset used in this study, a 5-fold cross-validation approach was employed to address potential issues that may arise from improper dataset partitioning. 5-fold cross-validation produces five sets of evaluation results, which are then averaged to produce a final result. The F1 score weighted by the label was used as the evaluation criterion. This advanced methodology produced highly accurate and reliable results.

For three selected rapport tasks, we experimented with one baseline model and seven ADVER models. All the machine learning models were trained with seven combinations of unimodal features (audio, visual, and linguistic features) to analyze the effectiveness of unimodal and multimodal features. The seven combinations of automatic feature sets were as follows:

(1) A: Model trained with acoustic features;
(2) V: Model trained with visual features;
(3) L: Model trained with linguistic features;

(4) A + V: Model trained with acoustic + visual features;
(5) A + L: Model trained with acoustic + linguistic features;
(6) V + L: Model trained with visual + linguistic features;
(7) ALL: Model trained with acoustic + visual + linguistic features.

6.2. Comparative Methods

6.2.1. Human Model

We prepared a human model using third-party annotations to evaluate the user rapport levels of dialogues. These third-party annotations were considered the outcomes of human perception. The average of the third-party annotation results was classified into high- and low-satisfaction categories based on a predefined threshold, and the F1 score was computed to evaluate the corresponding performance.

6.2.2. Instruction-Based LLM Model

We use GPT-4o to act as a dialogue expert and evaluate the overall dialogue by assessing the user's rapport. We carefully designed the prompt instruction by following [25] to output stable evaluation results. Specifically, we input the complete dialogue content and a carefully designed prompt into the model to obtain user rapport scores across different dimensions. The specific requirements of response are as follows:

Scoring: Assign a score for each rapport label of 0 (low) or 1 (high).
Explanation: Provide a brief justification for each score, explaining the reasoning behind the evaluation.

ChatGPT-4o is a cross-lingual model. Our testing revealed that English prompts can effectively evaluate Japanese dialogues. For clarity, the full prompt used to generate the evaluation is shown in Figure 6a, and Figure 6b presents an example of a response from ChatGPT-4o. As shown in the table, each score is accompanied by a rationale that explains the reasoning behind the model's assessment. In alignment with the human model, the F1 score was computed to evaluate the corresponding performance.

You are an expert dialogue evaluator. The following is a dialogue between a user and a dialogue system ;
{
User : .....
Agent : .....
User : .....
Agent : .....
.....
}
Based on this dialogue, please assign scores for the user's rapport using the following 18 labels:
•"well coordinated", boring", cooperative", harmonious", unsatisfying", uncomfortably paced", cold", awkward", engrossing", unfocused", involving", intense", friendly", active", positive", dull", worthwhile", and slow". .
Scoring Guidelines:
Each label should be scored with 0 (low) or 1 (high).

•The scores must be presented in the format: coordinated: 1, boring: 0, using commas to separate each score.
After providing the scores, include a brief explanation for each score to justify the evaluation.

(a) Prompt words

well coordinated: 1, boring: 0, cooperative: 1, harmonious: 1, unsatisfying: 0, uncomfortably paced: 0, cold: 0, awkward: 0, engrossing: 1, unfocused: 0, involving: 1, intense: 0, friendly: 1, active: 1, positive: 1, dull: 0, worthwhile: 1, slow: 0

well coordinated: 1 - The dialogue flows smoothly with questions and responses that are relevant and connected, indicating good coordination.
boring: 0 - The conversation covers a variety of interesting topics, such as food, travel, and technology, keeping it engaging.
cooperative: 1 - Both the user and the system contribute actively to the dialogue, maintaining cooperation throughout.
harmonious: 1 - There is a sense of agreement and mutual understanding in the dialogue, making it harmonious.
unsatisfying: 0 - The dialogue covers several topics and provides informative exchanges, making it satisfying.
uncomfortably paced: 0 - The pacing of the conversation seems comfortable and natural, without rushed segments.
cold: 0 - The dialogue has a warm and friendly tone, with personal anecdotes and shared experiences.
awkward: 0 - The interaction does not exhibit any awkwardness; it is smooth and coherent.
engrossing: 1 - The dialogue includes engaging topics and the user appears interested, making it engrossing.
unfocused: 0 - The conversation stays on track with specific topics, indicating focus and direction.
involving: 1 - The user is actively participating and sharing personal experiences, making the dialogue involving.
intense: 0 - The conversation is relaxed and friendly rather than intense.
friendly: 1 - The tone of the conversation is friendly, with expressions of interest and shared laughter.
active: 1 - Both the user and the system maintain an active participation throughout the dialogue.
positive: 1 - The dialogue maintains a positive tone, with expressions of enjoyment and interest.
dull: 0 - The variety of topics and the dynamic interaction prevent the dialogue from being dull.
worthwhile: 1 - The dialogue is meaningful and informative, making it a worthwhile interaction.
slow: 0 - The dialogue progresses at a steady and engaging pace, not slow.

(b) Example of Chagpt4-o response

Figure 6. Instruction Templates and Evaluation Responses for Instruction-Based LLM.

7. Results

Table 4 shows the F1 scores produced by the three unimodal [A, V, L] models across the three binary classification tasks. Table 5 presents the F1 scores obtained for the four multimodal feature sets [A + V, A + L, V + L, A + V + L] and the human model across the same three tasks. All the tasks are listed in eight sub rows, which consist of one baseline model and seven ADVER-based models (age, gender, extraversion, agreeableness, conscientiousness, neuroticism, and openness).

Table 4. Binary classification F1 score of different unimodal data for user rapport. (The bolded parts represent the modality that achieved the best result for each task.).

User Rapport	Personal Information	Unimodal		
		A	V	L
Well coordinated	Baseline	**0.526**	0.45	0.482
	Age	**0.685**	0.607	0.672
	Gender	**0.642**	0.449	0.583
	Extraversion	**0.613**	0.523	0.583
	Agreeableness	**0.764**	0.656	0.618
	Conscientiousness	**0.613**	0.5	0.552
	Neuroticism	0.505	0.549	**0.549**
	Openness	**0.662**	0.625	0.631
Awkward	Baseline	0.666	0.675	**0.7**
	Age	**0.709**	0.597	0.694
	Gender	0.648	0.665	**0.7**
	Extraversion	**0.719**	0.615	0.713
	Agreeableness	0.686	**0.725**	0.696
	Conscientiousness	0.64	0.642	**0.713**
	Neuroticism	0.608	0.587	0.587
	Openness	0.611	0.675	**0.714**
Engrossing	Baseline	0.659	0.671	**0.705**
	Age	0.658	0.67	**0.683**
	Gender	**0.708**	0.677	0.695
	Extraversion	0.684	0.681	**0.708**
	Agreeableness	0.699	0.711	**0.717**
	Conscientiousness	0.712	0.682	**0.72**
	Neuroticism	0.591	**0.681**	**0.681**
	Openness	0.693	0.679	**0.736**

Table 5. Binary classification F1 scores of different multimodal methods for user rapport. (The bolded parts represent the modality feature set that achieved the best result for each task.).

User Rapport	Personal Infromation	Multimodal				Human Model	Instruction-Based LLM Model
		A + V	A + L	V + L	ALL		
Well coordinated	Baseline	0.505	0.555	0.45	**0.611**	0.517	0.531
	Age	**0.723**	0.654	0.715	0.649		
	Gender	**0.643**	0.62	0.628	0.54		
	Extraversion	0.598	**0.663**	0.603	0.634		
	Agreeableness	0.676	0.704	0.669	**0.717**		
	Conscientiousness	0.558	0.55	**0.581**	0.576		
	Neuroticism	0.471	0.52	0.471	**0.531**		
	Openness	0.692	0.706	0.664	**0.722**		
Awkward	Baseline	0.682	0.655	**0.696**	0.673	0.607	0.365
	Age	0.684	0.674	**0.744**	0.711		
	Gender	0.678	0.619	**0.742**	0.729		
	Extraversion	0.679	0.662	**0.703**	0.675		
	Agreeableness	0.68	0.66	**0.719**	0.671		
	Conscientiousness	0.67	**0.703**	0.689	0.667		
	Neuroticism	0.649	0.593	0.649	**0.653**		
	Openness	0.641	**0.705**	0.698	0.675		
Engrossing	Baseline	**0.71**	0.69	0.665	0.655	0.571	0.540
	Age	0.647	**0.735**	0.713	0.666		
	Gender	0.683	**0.732**	0.674	0.665		
	Extraversion	0.62	0.693	**0.705**	0.685		
	Agreeableness	0.681	0.647	**0.729**	0.652		
	Conscientiousness	**0.725**	0.639	0.688	0.621		
	Neuroticism	**0.648**	0.632	0.648	0.627		
	Openness	**0.718**	0.653	0.691	0.698		

7.1. Efficacy of Adversarial Model for User Rapport Recognition (Answer to RQ1)

7.1.1. Unimodal Feature Comparison

Table 4 shows the three-task classification results obtained based on the unimodal features. The table shows the following.

- Well coordinated: The acoustic features achieved the best performance in the baseline. The age, gender, extraversion, agreeableness, and openness models improved upon the baseline, with the best result obtained by ADVER-Openness (0.526 to 0.764) with acoustic features.
- Awkward: The acoustic features yielded the best performance for the baseline (0.666). ADVER-Extraversion (0.719) achieved the best results, with a 0.53 improvement in the acoustic features. Among all the models, ADVER-Agreeableness (0.725) achieved the highest score with visual features.
- Engrossing: As shown in the table, the linguistic features produced the best results in the unimodal baseline (0.705). ADVER-Openness (0.736) achieved the highest score with linguistic features among all the adversarial models.

7.1.2. Multimodal Feature Comparison

Table 5 presents the classification results obtained for the three tasks based on multimodal features. The results were as follows.

- Well coordianted: In the multimodal experiments, the All feature set yielded the best performance in the baseline (0.611), representing an improvement of 0.085 over the best unimodal feature. ADVER-Gender (0.723) achieved the highest score with the A + V feature set, closely followed by ADVER-Openness (0.722) with the All feature set.
- Awkward: The V + L feature set produced the best performance for the baseline (0.696) among the multimodal feature sets. For the adversarial task model, the V + L feature set yielded the best result closely followed by ADVER-AGE (0.744), with ADVER-Gender (0.742).
- Engrossing: The A + V feature set yielded the best performance in the baseline (0.71), slightly improving upon that achieved with the unimodal features (0.705). In the adversarial experiments, the age, gender, agreeableness, conscientiousness, and openness models performed better. The A + L feature set yielded the best result with ADVER-AGE (0.735), closely followed by ADVER-GENDER with the A + L feature set (0.732).

Overall, the results of the adversarial experiments showed improvements in both unimodal and multimodal settings.

7.2. Impact of Demographic Data vs. Personality on User Rapport Recognition (Answer to RQ2)

To investigate the impact of personality and demographic data on user rapport in human-computer dialogue. We present the best F1 scores attained by the baseline and adversarial-based models with seven different multimodal feature sets, which are based on age, gender, and the Big Five personality traits in Table 6. Overall, we observed improvements in user rapport recognition impacts achieved with the ADVER-based methods. As indicated in Table 6, for demographic data (age and gender), the ADVER-based models generally demonstrated enhancements in all tasks. For the Big Five personality traits, most adversarial experiments showed improvements, except for the engrossed task, where ADVER-Extraversion slightly decreased (0.002) from the baseline result. Moreover, across all tasks, the performance of the ADVER-Neuroticism model decreased, suggesting a lack of useful information for predicting user attitudes within the neuroticism personality trait. Overall, ADVER-Agreeableness achieved the greatest improvement (0.153) in the well-coordinated task, whereas ADVER-Age achieved the greatest improvement (0.044) in the

awkward task. ADVER-Openness attained the greatest increase of 0.026 in the engrossed task. In summary, the adversarial learning models generally exhibited improvements in handling demographic data (age and gender). While overall performance enhancements were observed when addressing the Big Five personality traits, notable variations were exhibited across different personalities. Some personality traits, such as neuroticism, did not yield the expected improvements, whereas others, such as agreeableness and openness, demonstrated significant performance gains.

Table 6. Binary classification of user rapport results, "Diff" denotes the difference in F1-scores between the single task and adver_base models.

	Well Coordinated		Awkward		Engrossing	
	Best	**Diff**	**Best**	**Diff**	**Best**	**Diff**
Baseline	0.611	/	0.7	/	0.71	/
Age	0.723	0.112	0.744	0.044	0.735	0.025
Gender	0.643	0.032	0.742	0.042	0.732	0.022
Extraversion	0.663	0.052	0.719	0.019	0.708	−0.002
Agreeableness	0.764	0.153	0.725	0.025	0.729	0.019
Conscientiousness	0.613	0.002	0.713	0.013	0.725	0.015
Neuroticism	0.549	−0.062	0.653	−0.047	0.681	−0.029
Openness	0.722	0.111	0.714	0.014	0.736	0.026

7.3. Validating the Reliability of the Overall System (Answer to RQ3)

The human model and instruction-based LLM columns in Table 5 present the results for the human model and the instruction-based LLM model, respectively. In comparison with the human model, the instruction-based LLM model showed a +0.014 improvement for the "well coordinated" label. However, the instruction-based LLM model declined by a −0.031 for the "engrossing" label. Additionally, the performance on the "awkward" label was notably poor, with an F1 score of only 0.365, which is significantly lower than that of the other models. The human model and instruction-based LLM scores were considerably inferior to those of the machine learning baseline model for every task, and the proposed method further improved the baseline score for every task. More specifically, the proposed method significantly improved the well-coordinated task with +0.153 improvement and increased the scores achieved in the other rapport tasks to various degrees.

The entire proposed system not only outperformed the existing models in terms of performance but also demonstrated its reliability and effectiveness through empirical evidence, making it suitable for practical user rapport recognition applications.

8. Discussion

8.1. Feature Analysis

Combining the results shown in Tables 4 and 5, for the well-coordinated label, we found that the All feature set yielded the best performance in the baseline, which was consistent with the results in [23]. This finding aligns with the understanding that communication is a cooperative activity involving coordinated behaviors [43]. Furthermore, some studies have shown that dialogue participants spontaneously adjust their facial expressions, postures, pronunciation, and speech rates [44–46]. Among the adversarial models, ADVER-Agreeableness achieved the best result (0.764). The agreeableness personality traits was significantly related to coordination in dialogue. Individuals with high agreeableness are generally more cooperative and easier to work with, leading to more coordinated interactions in conversations. For the awkward label, the V + L feature set yielded the

best result in the baseline, which was consistent with [23,33]. We found that ADVER-Age achieved the best results in both the awkward and engrossing tasks, indicating that age significantly impacts rapport in conversations. Therefore, it is important to consider the user's age in the decision-making process of dialogue systems.

Neuroticism, as mentioned in Section 7.2, did not achieve the expected improvements in most of the tasks. Previous work [35] reported that neuroticism, even when annotated by humans, is significantly different from self-reported measures. Therefore, it is challenging for annotators to accurately judge a user's neuroticism based solely on a single dialogue. As [47] also indicates, predictions for neuroticism are less accurate than those for conscientiousness and extroversion. In our study, the poor performance of the neuroticism related model may be because the data in the dataset are based on casual conversations, where users typically engage in relaxed and humorous interactions. Such communication styles may not exhibit significant emotional fluctuations, especially with respect to neuroticism, which is typically associated with emotional instability and anxiety. As a result, the system may not capture enough emotional variation to identify these traits. In summary, this result may have occurred because our proposed model did not learn relevant information about neuroticism from a single dialogue, thus failing to achieve the expected improvements.

In future data collection efforts, we will focus on enabling the system to gather user interaction data across multiple conversational scenarios. For instance, discussions around stress, anxiety, or emotional topics can provide contrasting contexts that will help the system observe and identify personality traits such as neuroticism more effectively. By collecting data in a variety of scenarios, we can more accurately assess the manifestation of personality traits, thereby improving the system's accuracy and adaptability in recognizing personality characteristics, as well as enhancing its performance in identifying user rapport.

8.2. Effects of Adversarial Learning

To investigate whether the model genuinely learned features related to the task objective independently of users' personal information, we utilized an ADVER-based approach to recognize relevant users' personal information. Figure 7 presents the training loss of the "engrossing" label based on adversarial gender training with the A + L feature set. The graph illustrates the training losses for gender and the "engrossed" label across the different epochs in a single fold.

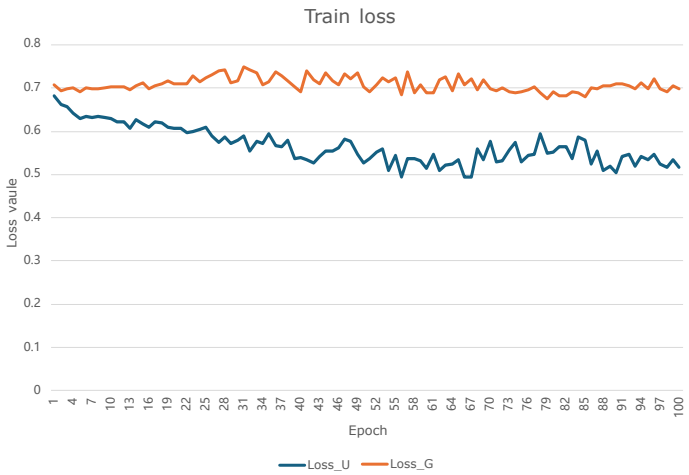

Figure 7. Train loss of "Engrossing" label based on gender adversarial training in the A + L feature set. Loss_U: train loss of the use rapport (engrossing), Loss_G: train loss of gender task.

The following is shown in Figure 7:

(1) Main task "engrossing" label loss: As training progressed, the performance achieved by the model in the main task improved, with the main task loss gradually decreasing until convergence was reached.
(2) The complexity of the adversarial gender loss: The adversarial task loss exhibited a more complex pattern, initially decreasing during the early stages of training. However, owing to the reversal effect of the GRL, the loss experienced fluctuations, reflecting the ongoing adaptations of the feature extractor to the requirements of domain-adversarial training.

To gain further insights into the impact of the engrossing labels, we separately compiled the results of different models for the engrossing labels. Figure 8 presents the confusion matrix depicting these results. This figure shows that the human model achieved the best performance in high engrossing at 53.6%. Instruction-based LLM performed poorly in both the low-engrossing and high-engrossing categories. Additionally, both the human model and the instruction-based LLM yielded incorrect low-level recognition results (misclassifying true low engrossing as high engrossing), accounting for 22.8% and 22.4% of the total samples, respectively, which was significantly higher than the baseline and ADVER-Gender (12.8% and 12.0%). The poor performance of the human model and instruction-based LLM may be attributed to the lack of domain-specific training and fine-tuning, which hinders their ability to capture task-related information effectively. Additionally, for the instruction-based LLM model, the biases in pre-trained models and their limited understanding of the entire dialogue context also affect their performance. Compared with the baseline, ADVER-Gender demonstrates improvements in both high and low-engrossing labels, with increases of 3.2% and 0.8%, respectively.

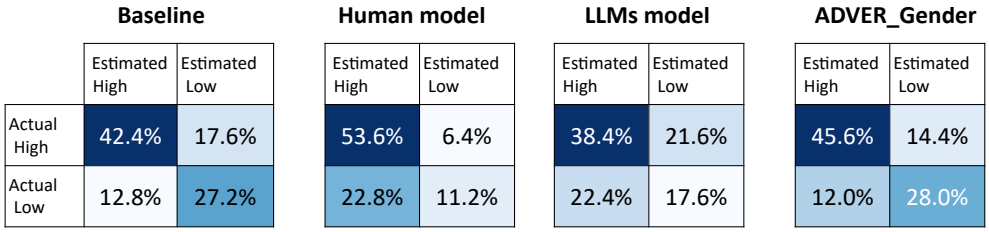

Figure 8. The confusion matrix of baseline, human model, instruction-based LLM model, and ADVER-Gender model for the engrossing label in the A + L feature set. (Darker colors represent higher values, and lighter colors represent lower values.).

In conclusion, these results demonstrate that the model has indeed learned gender-independent features while also enhancing the recognition performance of the labels.

9. Conclusions

In this work, we first investigated the relationship between users' personal information and their rapport. We found that age, gender, and personality differences do influence user rapport. To address this influence, we proposed a domain-adversarial model that reduces the impact of user trials by learning adversarial features that are unrelated to users' personal information. The results indicate that our proposed adversarial learning model achieved a significant performance improvement. Moreover, our system consistently demonstrated its superior performance to the annotations by the human and instruction-based LLM models, thereby confirming the reliability and effectiveness of our system. In addition to the limitations imposed by the size of the utilized dataset, our model still has room for improvement. We continue to collect relevant data across multiple conversational scenarios and closely monitor the release of new datasets suitable for our research. As the data at the

sentence level are sufficient, we will try to conduct analyses at the sentence level to explore the impact of users' personal information on users during human-computer interactions.

Author Contributions: Conceptualization, W.W., K.K., S.L. and S.O.; methodology, W.W., S.L. and S.O.; validation, W.W., S.L. and S.O.; Formal analysis, W.W., C.O.M., S.L. and S.O.; writing—original draft preparation, W.W. and X.L.; writing—review and editing, W.W., X.L., S.L., C.O.M., K.K. and S.O.; visualization, W.W., S.L. and S.O.; supervision, S.O., K.K., S.L. and C.O.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by JSPS KAKENHI (22H04860, 22K21304, 22H00536, 23H03506), and JST AIP Trilateral AI Research, Japan (JPMJCR20G6).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study are included in this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Huang, M.; Zhu, X.; Gao, J. Challenges in building intelligent open-domain dialog systems. *ACM Trans. Inf. Syst. (TOIS)* **2020**, *38*, 21. [CrossRef]
2. Ling, Y.; Liang, Z.; Wang, T.; Cai, F.; Chen, H. Sequential or jumping: Context-adaptive response generation for open-domain dialogue systems. *Appl. Intell.* **2022**, *53*, 11251–11266. [CrossRef]
3. Young, T.; Xing, F.; Pandelea, V.; Ni, J.; Cambria, E. Fusing task-oriented and open-domain dialogues in conversational agents. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 11622–11629.
4. Iizuka, S.; Mochizuki, S.; Ohashi, A.; Yamashita, S.; Guo, A.; Higashinaka, R. Clarifying the Dialogue-Level Performance of GPT-3.5 and GPT-4 in Task-Oriented and Non-Task-Oriented Dialogue Systems. In Proceedings of the AAAI Symposium Series, Burlingame, CA, USA, 27–29 March 2023; Volume 2, pp. 182–186.
5. Degnen, L.; Rosenthal, R. The nature of rapport and its nonverval correlates. *Psychol. Inq.* **1990**, *1*, 285–293. [CrossRef] [PubMed]
6. Müller, P.; Huang, M.X.; Bulling, A. Detecting low rapport during natural interactions in small groups from non-verbal behaviour. In Proceedings of the 23rd International Conference on Intelligent User Interfaces, Sydney, Australia, 27–31 March 2018; pp. 153–164.
7. Abbe, A.; Brandon, S.E. The role of rapport in investigative interviewing: A review. *J. Investig. Psychol. Offender Profiling* **2013**, *10*, 237–249. [CrossRef]
8. Hayashi, T.; Mawalim, C.O.; Ishii, R.; Morikawa, A.; Fukayama, A.; Nakamura, T.; Okada, S. A ranking model for evaluation of conversation partners based on rapport levels. *IEEE Access* **2023**, *11*, 73024–73035. [CrossRef]
9. Cerekovic, A.; Aran, O.; Gatica-Perez, D. Rapport with Virtual Agents: What Do Human Social Cues and Personality Explain? *IEEE Trans. Affect. Comput.* **2017**, *8*, 382–395. [CrossRef]
10. Shiota, M.N.; Keltner, D.; John, O.P. Positive emotion dispositions differentially associated with Big Five personality and attachment style. *J. Posit. Psychol.* **2006**, *1*, 61–71. [CrossRef]
11. Penley, J.A.; Tomaka, J. Associations among the Big Five, emotional responses, and coping with acute stress. *Personal. Individ. Differ.* **2002**, *32*, 1215–1228. [CrossRef]
12. Wei, Z.; Liu, Q.; Peng, B.; Tou, H.; Chen, T.; Huang, X.J.; Wong, K.F.; Dai, X. Task-oriented dialogue system for automatic diagnosis. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Volume 2, pp. 201–207, Short Papers.
13. Zhang, Z.; Takanobu, R.; Zhu, Q.; Huang, M.; Zhu, X. Recent advances and challenges in task-oriented dialog systems. *Sci. China Technol. Sci.* **2020**, *63*, 2011–2027. [CrossRef]
14. Zhao, M.; Wang, L.; Jiang, Z.; Li, R.; Lu, X.; Hu, Z. Multi-task learning with graph attention networks for multi-domain task-oriented dialogue systems. *Knowl.-Based Syst.* **2023**, *259*, 110069. [CrossRef]
15. Kobyashi, S.; Hagiwara, M. Non-task-oriented dialogue system considering user's preference and human relations. *Trans. Jpn. Soc. Artif. Intell.* **2016**, *31*. [CrossRef]
16. Inaba, M.; Iwata, N.; Toriumi, F.; Hirayama, T.; Enokibori, Y.; Takahashi, K.; Mase, K. Constructing a non-task-oriented dialogue agent using statistical response method and gamification. In Proceedings of the International Conference on Agents and Artificial Intelligence, Angers, France, 6–8 March 2014; SCITEPRESS: Setubal, Portugal, 2014; Volume 2, pp. 14–21.

17. Jekosch, U. *Voice and Speech Quality Perception: Assessment and Evaluation*; Springer Science & Business Media: Heidelberg, Germany, 2006.
18. Engelbrecht, K.P.; Gödde, F.; Hartard, F.; Ketabdar, H.; Möller, S. Modeling user satisfaction with hidden Markov models. In Proceedings of the SIGDIAL 2009 Conference, London, UK, 11–12 September 2009; pp. 170–177.
19. Ultes, S. Improving interaction quality estimation with BiLSTMs and the impact on dialogue policy learning. *arXiv* **2020**, arXiv:2001.07615.
20. Hirano, Y.; Okada, S.; Nishimoto, H.; Komatani, K. Multitask prediction of exchange-level annotations for multimodal dialogue systems. In Proceedings of the 2019 International Conference on Multimodal Interaction, Suzhou, China, 14–18 October 2019; pp. 85–94.
21. Kim, T.E.; Lipani, A. A multi-task based neural model to simulate users in goal oriented dialogue systems. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 11–15 July 2022; pp. 2115–2119.
22. Schmitt, A.; Ultes, S. Interaction quality: Assessing the quality of ongoing spoken dialog interaction by experts—And how it relates to user satisfaction. *Speech Commun.* **2015**, *74*, 12–36. [CrossRef]
23. Wei, W.; Li, S.; Okada, S.; Komatani, K. Multimodal user satisfaction recognition for non-task oriented dialogue systems. In Proceedings of the 2021 International Conference on Multimodal Interaction, Montréal, QC, Canada, 18–22 October 2021; pp. 586–594.
24. Bodigutla, P.K.; Tiwari, A.; Vargas, J.V.; Polymenakos, L.; Matsoukas, S. Joint turn and dialogue-level user satisfaction estimation on multi-domain conversations. *arXiv* **2020**, arXiv:2010.02495.
25. Mendonça, J.; Trancoso, I.; Lavie, A. Soda-Eval: Open-Domain Dialogue Evaluation in the age of LLMs. *arXiv* **2024**, arXiv:2408.10902.
26. Mendonça, J.; Trancoso, I.; Lavie, A. ECoh: Turn-level Coherence Evaluation for Multilingual Dialogues. *arXiv* **2024**, arXiv:2407.11660.
27. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 46595–46623.
28. Laskar, M.T.R.; Alqahtani, S.; Bari, M.S.; Rahman, M.; Khan, M.A.M.; Khan, H.; Jahan, I.; Bhuiyan, A.; Tan, C.W.; Parvez, M.R.; et al. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 12–16 November 2024; pp. 13785–13816.
29. Meng, Z.; Li, J.; Gong, Y. Adversarial speaker adaptation. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 5721–5725.
30. Gao, Y.; Okada, S.; Wang, L.; Liu, J.; Dang, J. Domain-invariant feature learning for cross corpus speech emotion recognition. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 6427–6431.
31. Komatani, K.; Okada, S. Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels. In Proceedings of the 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII), Nara, Japan, 28 September–1 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–8.
32. Katada, S.; Okada, S.; Komatani, K. Transformer-based physiological feature learning for multimodal analysis of self-reported sentiment. In Proceedings of the 2022 International Conference on Multimodal Interaction, Bengaluru, India, 7–11 November 2022; pp. 349–358.
33. Wei, W.; Li, S.; Okada, S. Investigating the relationship between dialogue and exchange-level impression. In Proceedings of the 2022 International Conference on Multimodal Interaction, Bengaluru, India, 7–11 November 2022; pp. 359–367.
34. Lee, A.; Oura, K.; Tokuda, K. MMDAgent—A fully open-source toolkit for voice interaction systems. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 8382–8385.
35. Komatani, K.; Takeda, R.; Okada, S. Analyzing differences in subjective annotations by participants and third-party annotators in multimodal dialogue corpus. In Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Prague, Czechia, 11–15 September 2023; pp. 104–113.
36. 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. Personal. Soc. Psychol.* **1996**, *71*, 110. [CrossRef]
37. Oshio, A.; Shingo, A.; Cutrone, P. Development, reliability, and validity of the Japanese version of Ten Item Personality Inventory (TIPI-J). *Jpn. J. Personal./Pasonariti Kenkyu* **2012**, *21*, 40–52. [CrossRef]

38. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affect. Comput.* **2015**, *7*, 190–202. [CrossRef]
39. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
40. Baltrusaitis, T.; Zadeh, A.; Lim, Y.C.; Morency, L.P. Openface 2.0: Facial behavior analysis toolkit. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 59–66.
41. Cohn, J.F.; Ambadar, Z.; Ekman, P. Observer-based measurement of facial expression with the Facial Action Coding System. *Handb. Emot. Elicitation Assess.* **2007**, *1*, 203–221.
42. Ganin, Y.; Lempitsky, V. Unsupervised Domain Adaptation by Backpropagation. *arXiv* **2015**. [CrossRef]
43. Latif, N.; Barbosa, A.V.; Vatiokiotis-Bateson, E.; Castelhano, M.S.; Munhall, K. Movement coordination during conversation. *PLoS ONE* **2014**, *9*, e105036. [CrossRef] [PubMed]
44. Cappella, J.N.; Planalp, S. Talk and silence sequences in informal conversations III: Interspeaker influence. *Hum. Commun. Res.* **1981**, *7*, 117–132. [CrossRef]
45. McHugo, G.J.; Lanzetta, J.T.; Sullivan, D.G.; Masters, R.D.; Englis, B.G. Emotional reactions to a political leader's expressive displays. *J. Personal. Soc. Psychol.* **1985**, *49*, 1513. [CrossRef]
46. Pardo, J.S. On phonetic convergence during conversational interaction. *J. Acoust. Soc. Am.* **2006**, *119*, 2382–2393. [CrossRef] [PubMed]
47. Ivanov, A.V.; Riccardi, G.; Sporka, A.J.; Franc, J. Recognition of personality traits from human spoken conversations. In Proceedings of the Twelfth annual Conference of the International Speech Communication Association, Florence, Italy, 27–31 August 2011.