IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 9, 2020, accepted December 19, 2020, date of publication January 13, 2021, date of current version January 22, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3051526

Estimating Subjective Argument Quality Aspects From Social Signals in Argumentative Dialogue Systems

NIKLAS RACH[1,2], YUKI MATSUDA[2,3], (Member, IEEE), STEFAN ULTES[4], WOLFGANG MINKER[1], AND KEIICHI YASUMOTO[2], (Member, IEEE)

[1]Institute of Communications Engineering, Ulm University, 89081 Ulm, Germany
[2]Graduate School of Science and Technology, Nara Institute of Science and Technology, Ikoma 630-0192, Japan
[3]JST PRESTO, Tokyo 102-0076, Japan
[4]Mercedes-Benz Research and Development, 71059 Sindelfingen, Germany

Corresponding author: Niklas Rach (niklas.rach@uni-ulm.de)

Parts of this work have been funded by the Deutsche Forschungsgemeinschaft (DFG) within the project ''How to Win Arguments - Empowering Virtual Agents to Improve their Persuasiveness'', Grant Number 376696351, as part of the Priority Program ''Robust Argumentation Machines (RATIO)'' (SPP-1999).

ABSTRACT Information about a subjective user opinion towards an argument is crucial for argumentative systems in order to present appropriate content and adapt their behaviour to the individual user. However, requesting explicit feedback regarding the discussed arguments is often impractical and can hinder the interaction. To address this issue, we investigate the automatic recognition of user opinions towards arguments that are presented by means of a virtual avatar from social signals. We focus on two different user opinion categories (*convincing* and *interesting*) and two different types of social signals (facial expressions and eye movement). The recognition is addressed as a supervised learning problem and realized using the argument search evaluation data discussed in previous work. The overall performance is compared to a human annotation on a subset of the collected data. The results show that the machine learning performance is similar to human performance in both recognition tasks.

INDEX TERMS Computational argumentation, argument quality estimation, argumentative dialogue systems, social signal extraction, machine learning.

I. INTRODUCTION

In recent years, argumentation has gained a lot of interest as a domain for dialogue systems, chatbots and conversational agents. The tasks addressed by such argumentative systems range from full scale debates with a human counterpart[1] over persuasion [1], [2] to customer support [3] and opinion building [4]. Despite the different natures of all these tasks, the ability to complete them depends on the availability of arguments and their quality. Although the assessment of argument quality has been discussed extensively (see for example [5] for an overview), the perception of an argument by a user remains subjective [6]. Especially for tasks that require an adaptation to the individual user, the personal opinion of this user is more relevant than an objective quality score. At the same time, explicit user feedback as for example discussed in [7] is often impractical and can hinder the interaction. In such cases, an automatic assessment of user opinions and perceptions of arguments is required.

To address this issue, we present an approach to estimate subjective quality aspects of arguments presented by an argumentative dialogue system from social signals shown by the user during the presentation. The estimated aspects are the users *interest* in an individual argument as well as its perceived *convincingness*. Both aspects were used in [6] to assess the suitability of arguments retrieved by means of argument search engines for completing argumentative tasks. The assessment was done through explicit ratings by users given after the argument was presented by means of synthetic speech and by a virtual avatar. In the present work, we build upon the therein introduced approaches and results by using the explicit user ratings as labels for each argument and estimate them from social signals shown by the corresponding

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Liu.

[1] https://www.research.ibm.com/artificial-intelligence/project-debater/

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

participant during the interaction. The proposed method is along the line of affective computing approaches that aim at estimating subjective information like interest [8] or agreement [9] from social cues. Whereas some approaches utilize a mapping of the observed cues to a psychological model of emotions and derive the final label from the recognized emotion (as for example in [10]), we directly estimate the quality labels from the observed cues. The herein utilized social signals include eye movement and facial expressions extracted from video recordings of the interactions. In addition, we perform a human annotation for both aspects on a (randomized) representative test set in order to compare the automatic estimation to human performance. The comparison shows similar performances with a slight advantage of the human annotation in the *convincing* task whereas the automatic estimation performs slightly better in the *interesting* task. In summary, the contributions of this work are:

- Providing the (to the best of our knowledge) first recognition of argument quality aspects based on individual social signals.
- Discussion of the contributing features and their importance.
- Comparison of this estimation to human performance.

The remainder of this paper is as follows: Section II provides an overview of related work from the fields of computational argumentation and affective computing. Section III introduces both investigated quality aspects, including their motivation in the context of task success of dialogue systems. In Section IV, the utilized dialogue system is introduced whereas Section V covers the data collection. A discussion of the data statistics and feature computation is provided in Section VI. The machine learning setup, including parameter selection and results, is presented in Section VII, followed by the discussion of the human annotation in Section VIII. Finally, we discuss our findings in Section IX and close the paper with a conclusion and an outlook on future research directions in Section X.

II. RELATED WORK

Throughout this section, we give an overview of related work from two different perspectives. First, we discuss recent work on argument quality and recall existing approaches to assess it. Secondly, we provide an overview of related approaches from the field of affective computing with an emphasis on the two tasks of agreement and interest recognition.

A. ARGUMENT QUALITY

For computational argumentation, Wachsmuth *et al.* [5] presented a unified taxonomy for the theoretical assessment of argument quality in different sources and for different argument granularities. They divided argument quality into the broad categories of *logical*, *rhetorical* and *dialectical* quality and introduced 15 fine-grained sub-dimensions as well as a corpus annotated with these dimensions. Based on this theoretical approach, an annotated multi-domain corpus for argument quality assessment was introduced and models to estimate the quality scores from the annotated arguments were investigated [11]. The authors in [12] discussed an approach to assess the convincingness of arguments in which arguments were rated in direct comparison to each other in a crowd-sourcing experiment. The explanations provided by the annotators for their ratings were then used to infer argument quality properties. In addition, a corpus for the pairwise comparison of the convincingness of evidence was introduced in [13]. The correlations between the theoretical and the comparative crowd-sourcing approach were also investigated [14]. The authors found that a lot of reasons provided for the annotations in the crowd-sourcing setup could be matched to theoretical quality dimensions and that the comparative annotations correlate with theory-based absolute ratings. The overall quality of single arguments as well as argument pairs was discussed by Toledo *et al.* [15] together with automatized approaches for argument ranking and argument-pair classification. A comparison of the single argument and the argument pair annotation showed that the results of both approaches are mostly consistent and the authors suggested to use pair-wise approaches mainly for argument pairs with a low difference in the individual rating. In addition, a large corpus with binary annotations regarding the quality of arguments and their stance was introduced and methods to derive argument quality scores from the labels were investigated alongside a prediction model for argument quality [16]. Finally, Potthast *et al.* [17] utilized expert ratings of the above mentioned categories *logical*, *rhetorical* and *dialectical* quality to compare different retrieval approaches for argument search in combination with the information retrieval notion of relevance. Despite the remarkable results of these works, the aim of estimating the subjective user perception of arguments was (to the best of our knowledge) not addressed so far.

B. AFFECTIVE COMPUTING

On the affective computing side, several approaches that bear similarity to the herein discussed tasks exist. In particular, the automatic recognition of (dis)agreement is closely related to the herein considered *convincing* task. Following the notion of Poggi *et al.* [18], for two persons A and B, *"[...] B agrees with A when B assumes that his/her opinion is the same, similar or in any case congruent (in the same line, not conflicting)"*. Consequently, we can say that if B is convinced by an argument of A, he/she also agrees with A (on the topic of the argument). On the other hand, agreement does not necessarily implicate that the agreeing person is convinced. For example, A and B can agree on a premise but draw different conclusions from it, therefore establishing contradicting opinions and arguments. In addition, social signals associated with agreement can also be shown out of politeness or as a sign of attention (apparent agreement, [18]). Consequently, we can say that if a person A is convinced he/she is likely to show signs of agreement but signs of agreement do not necessarily implicate a convincing argument.

From a technical perspective, a lot of different approaches to recognizing (dis)agreement were introduced and overviews are provided for example in [9], [19]. The authors distinguish three types of (dis)agreement – direct speakers (dis)agreement, indirect speakers (dis)agreement and nonverbal listeners (dis)agreement – and discuss the corresponding indicating cues. Besides verbal indicators gestures, postures, facial expressions and eye movements are listed as frequent cues. Regarding the herein considered task, the non-verbal cues are of more interest as the participants assumed a passive and quiet role. An approach to (dis)agreement recognition that utilizes only non-verbal cues was presented in [20] and serves as a reference point for the comparison of convincingness recognition and (dis)agreement recognition.

For the second task considered throughout this work, namely the recognition of the participants' interest in the presented arguments, multiple related approaches were introduced. An approach closely related to emotion recognition that estimates the level of interest from facial expressions was presented in [10]. Moreover, Schuller *et al.* [21] recorded a multimodal corpus of human-human conversation and annotated the level of interest. Afterwards, different features including facial expressions, eye movements as well as linguistic and acoustic features were utilized to estimate the annotated labels. The collected data was also used in [22] to estimate utterance wise interest levels of users from acoustic and lexical features. In [23], the user interest in movie trailers was assessed and estimated from facial expressions as well as biological signals, whereas Hirayama *et al.* [24] estimated user interest in displayed content from eye movements. For the case of human-machine dialogue, facial expressions alongside prosodic features related to the user's voice were used to determine the turn-wise interest of the human user in the discussed topic [8]. Finally, a cognitive model of user interest in different objects for the application in smart environments was introduced in [25]. However, the interest in individual arguments as investigated throughout this work was, to the best of our knowledge, not considered so far.

III. EVALUATION CATEGORIES

The herein discussed approach utilizes data collected in an experiment that assessed the usability of argument search engines in argumentative dialogue systems. Throughout this section, we first recall the corresponding evaluation categories and subsequently discuss the selection of the two categories that are estimated in the present work. In the following, we denote a single search result from an argument search engine as *argument* and its polarity towards the overall topic as *stance* (support/PRO or attack/CON).

A. ORIGINAL CATEGORIES

Argument search engines in general aim at retrieving arguments regarding a given topic from multiple sources by following different paradigms [26]. The argument search engines in the original setup were therefore used to retrieve arguments and their stance regarding four different topics. The assessment of their usability in dialogue systems was approached by evaluating the following two broad aspects:

a) The structural properties of the retrieved arguments that are influenced by the different technological approaches (i.e., identification of arguments and stance) of the search engines.
b) The suitability of the retrieved arguments for the different tasks of an argumentative system.

Aspect a) was addressed directly through *argument quality* criteria that are strongly influenced by the technological differences between the search engines. The corresponding categories were *comprehensible* (Does the argument make sense by itself?) and *related* (Is the presented argument related to the topic and is the presented relation correct?).

For aspect b), the general notion of *task success* [27] as a common approach to assess task-oriented dialogue systems was used. Since the success of an argumentative dialogue system often depends on the individual user and is therefore hard to measure objectively, properties of the retrieved arguments that facilitate the completion of possible tasks were identified and assessed in separate categories. Based on the types of argumentative dialogue [28], the different tasks of argumentative dialogue systems were broadly divided into competitive and cooperative tasks. In competitive tasks like persuasion or negotiation, the overall goal is to *convince* the opposite side of for example a certain point of view (persuasion) or to accept a specific offer (negotiation). Therefore, the relevant property of the utilized arguments is their overall likeliness to convince the opponent, in short, their *convincingness*. The respective evaluation category was hence *convincing* (Does the argument convince me?) for competitive setups.

Cooperative setups (for example deliberation) on the other hand aim for a mutual solution of an issue by exchanging arguments that contribute to this task. In contrast to competitive setups, the goal of the involved parties is not to convince the other participants of a certain point of view but to find the best common ground. Therefore, the suitability of an argument for these tasks depends on its ability to contribute to this solution. However, in an argumentative application, this common ground should satisfy the user's needs and hence depends on the user perspective on the presented arguments. Consequently, this property was condensed in the question if an argument is *interesting* for the user given the discussed topic and assessed by means of a category with the same name.

B. ESTIMATED CATEGORIES

Throughout this work, we focus on the task-related categories *convincing* and *interesting* and estimate the respective user rating from social signals. This choice is due to several reasons: First of all, for both categories, similar approaches in the field of affective computing exist, namely the estimation of (dis)agreement and interest in other domains. Therefore, it is likely that users show distinct non-verbal cues that can be used to estimate the corresponding rating. Secondly, the categories *related* and *comprehensible* are meant to assess technical differences between argument search approaches and are less likely to be influenced by their presentation through the system. In contrast, the user perception of the categories *convincing* and *interesting* can also be changed by different presentations of the argument. In the case of the *convincing* category, the presentation of the argument was discussed as a quality (sub)dimension in [5] and approaches to increase the convincingness of an argumentative dialogue system by adapting for example the behaviour of the presenting virtual avatar were already introduced [7]. Similarly, the user's interest and his or her engagement in the interaction with a conversational system was shown to be influenced by the system's presentation of the content [29]. Consequently, the information whether or not a user is convinced by or interested in a presented argument can be used for adapting the presentation of following arguments whereas the information about a correct/incorrect relation and the comprehensibility of an argument cannot be transferred to the next one. Also, since the perception of all categories is highly subjective, the information about the relation and the comprehensibility cannot necessarily be used in following interactions with different users. In summary, we focus on the estimation of the categories *convincing* and *interesting* because we expect indicating signals from users based on existing work and because the estimated information can directly be used for system adaptation.

IV. SYSTEM

The dialogue system used for the data collection allows users to apply the evaluation criteria discussed throughout Section III-A in an intuitive way and during the ongoing interaction. The system was designed specifically for the evaluation task and allows the user to select his or her ratings as a direct response to the system utterance. The rating for each category can be given once for each argument and cannot be changed. In addition, the user is able to start the conversation, request the next argument, go to the previous one and repeat the latest utterance that includes an argument. If requested, the system selects the next argument randomly from the pool of available ones but each argument can occur only once during the interaction. The overall interaction is stopped by the system after a fixed time to ensure the same conditions for each user. It is important to note that the system requirements regarding the utilized arguments are as liberal as possible in order to enable the comparison of multiple different search approaches. The only information that has to be provided is the content of the argument as well as a notion of the respective stance towards the main topic. In the scope of the herein discussed experiment, the system had access to arguments collected with three different argument search engines, namely ArgumenText [30], args.me [31] and a baseline system [6]. Whereas ArgumenText and the baseline system retrieve arguments on a sentential level that can directly be used by the evaluation system, args.me provides more extensive arguments (consisting of one post on a debating website) which had to be shortened to a reasonable length (60 words) for the herein discussed scenario.

A. INTERFACE

The interface is adapted from the systems introduced in [32] and [7]. It is based on the Charamel™ avatar[2] which presents the system utterance via synthetic speech by utilizing Nuance TTS and Amazon Polly Voices.[3] Besides the avatar, the interface also includes buttons for the ratings in each category as well as the remaining user options (repeat, next, previous, start). If an option is not available in the current state of the interaction (e.g. if a rating was already given for the current argument), this is indicated by the appearance of the respective buttons. A screenshot of the interface including buttons and avatar is shown in Figure 1.

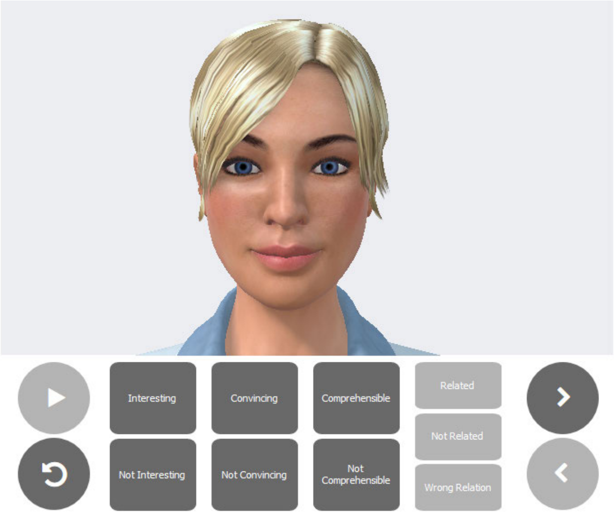

FIGURE 1. Screenshot of the system, including avatar and navigation buttons.

B. NATURAL LANGUAGE GENERATION

The system utterance is generated by a modified version of the template-based Natural Language Generation (NLG) used in [32] that receives the list of retrieved arguments from a certain argument search approach as input. Each argument is presented with an explicit notification of the assigned stance regarding the discussed topic. The specific formulation is selected randomly from a list of available options. In addition, the system starts each interaction by giving a topic-specific introduction including a short repetition of the task and the general claim *<TOPIC> is/are good*. Moreover, statements to notify the user when no further arguments are left and when the time for the discussed topic is over are also included. Table 1 shows the transcript of an interaction with the system, including the different utterance types of the dialogue system, the corresponding system utterance and exemplary

[2] https://www.charamel.com/competence/avatare

[3] https://docs.aws.amazon.com/polly/latest/dg/voicelist.html

TABLE 1. Dialogue transcript for the topic *Veganism*, including excerpts of the introduction and the closing statement, three arguments and the corresponding exemplary user ratings.

	System Utterance	Ratings
Intro	[...] The topic I want to discuss first is: *Veganism is good.*	-
Arg$_1$	In contradiction to my initiating claim I found the following opinion: *Eating animals is in our nature, therefore it cannot be wrong.*	interesting not convincing comprehensible related
Arg$_2$	Okay, let's continue with a new aspect. The next argument is a supporting one. The content is: *There is a nice vegan restaurant, just around the corner.*	not interesting not convincing comprehensible not related
Arg$_3$	All right, let me see what else I found. Next in line is a support argument with the following content: *A key motivation for veganism is to prevent the exploitation of animals.*	interesting convincing comprehensible related
End	Thank you for your ratings. Unfortunately, we reached the time limit for this topic. [...]	-

user ratings. The arguments in this example were created manually for demonstration.

V. DATA COLLECTION

The evaluation experiment was conducted with 19 international students (14 male, 5 female) at Nara Institute of Science and Technology (NAIST) and approved by the Ethics Review Committee of NAIST. Before the experiment started, participants received instructions about the overall purpose of the experiment, the system interface and the meaning of the categories. In addition, a short test trial with a separate topic was offered to exemplify the procedure. During the experiment, participants were seated on a chair, facing a 27inch display that shows the system in full-screen mode. A camera was placed on top of the display in order to record the user reactions throughout the experiment. A picture of the complete setup is shown in Figure 2. Participants listened to arguments presented by the virtual avatar and gave quality ratings in the four categories *convincing*, *interesting*, *comprehensible* and *related*. Each participant listened to arguments related to the four topics *nuclear energy*, *animal testing*, *self-driving cars* and *capital punishment* with a time limit of five minutes per topic after which the system stopped the rating process and introduced the next topic. The arguments for the first three topics were retrieved with one of the three argument search engines, whereas the arguments for the topic *capital punishment* included arguments collected with both ArguMenText and args.me and were the same for each participant. The ratings for the first three topics were used to compare the utilized argument search engines [6] whereas the ratings for the fourth topic were used to analyse the agreement between the participants in each category by means of Krippendorff's α [33]. The found agreement was low for all categories, with the highest alpha of 0.15 in the *comprehensible* category which emphasises the subjective nature of the task that is the main motivation for the herein discussed approach.

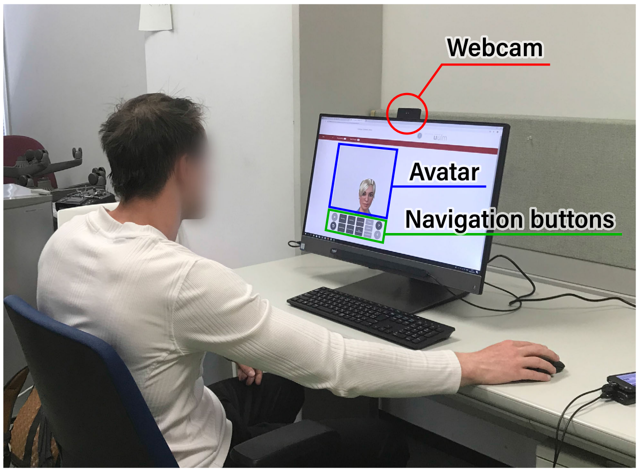

FIGURE 2. Experimental setup.

Besides the rating of the arguments, participants answered two questionnaires (one before and one after the interaction). In the first survey, the participants' stance on the discussed topics and their proficiency regarding dialogue systems were assessed, whereas the questions after the interaction were concerned with the understandability of the system with an emphasis on synthetic speech and language skills of the participants. In case a participant reported substantial difficulties with the language, the corresponding ratings were excluded from the data set (this occurred only once). Although several participants reported that they were irritated sometimes by the synthetic speech, no participant reported that he or she was not able to understand the presented content as can be seen in Figure 3. Consequently, we assume that the reactions shown by the users are mainly due to the presented content, although the reported irritation by the synthetic voice may introduce some noise for the recognition of the user opinion.

VI. DATA STATISTICS AND FEATURE EXTRACTION

The experiment resulted in a total of 1263 ratings in the two categories *interesting* (653) and *convincing* (610) and approximately 6 hours of recorded interaction. The difference in the number of ratings for both categories is due to the fact that participants were able to skip a rating in one or more categories if they were undecided. The splitting into the two categories and the respective class balance (positive and negative response) is shown in Table 2. We see that in the *convincing* case, the distribution of positive and negative ratings is almost balanced, whereas in the *interesting* case the positive ratings outweigh the negative ones.

For the classification approach, we assume that the non-verbal signals that correspond to the ratings are shown

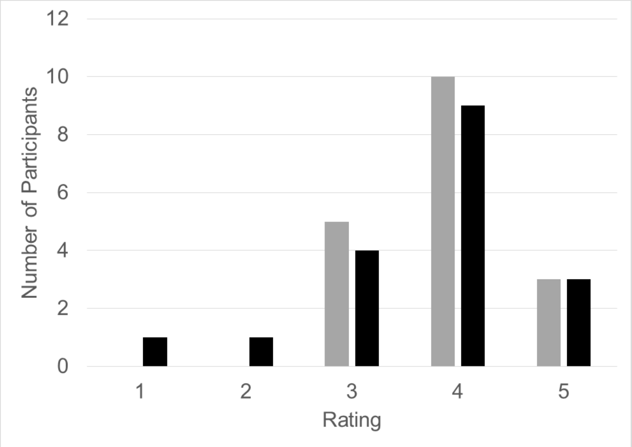

FIGURE 3. Responses on a five-point Likert scale from completely disagree (1) to fully agree (5) for the statements *The synthetic speech was easy to understand* (black bars) and *All in all I had no problems understanding the system utterances* (grey bars).

TABLE 2. Statistics of the recorded data for the two categories *interesting* and *convincing*.

	Convincing	Interesting
Positive	287	506
Negative	323	147
Total	610	653
Ratio	0.47	0.77

TABLE 3. Action units extracted by the OpenFace toolbox [34], including number and description.

Action Unit	Description
AU01	Inner Brow Raiser
AU02	Outer Brow Raiser
AU04	Brow Lowerer
AU05	Upper Lid Raiser
AU06	Cheek Raiser
AU07	Lid Tightener
AU09	Nose Wrinkler
AU10	Upper Lid Raiser
AU12	Lip Corner Puller
AU14	Dimpler
AU15	Lip Corner Depressor
AU17	Chin Raiser
AU20	Lip Stretcher
AU23	Lip Tightener
AU25	Lips Part
AU26	Jaw Drop
AU45	Blink

during or directly after the presentation of the argument since participants are focused solely on the content of the arguments during this time. Consequently, we investigate the time window between the user request for a new argument and his or her first rating for it. If the user listens repeatedly to the argument, the whole time window until the first rating is included. We call a pair of rating and time window a session and denote the time window corresponding to a specific session i with w_i.

Due to the passive role of the participants in our setup, we assume that the most informative cues are shown on the participants' face and focus on social signals related to facial expressions and eye movement. To derive a set of meaningful features for the classification, we extract the intensity of the 17 facial action units (AUs) shown in Table 3 and 8 eye movement values (x, y, z and angle for both eyes). The extraction is done by means of the OpenFace toolbox [34], resulting in a data point $q_t = (AU_{1-17,t}, e_{1-8,t})$ with the dimension $d = 25$ for each video frame. Based on all data points $\{q_t\}$ within a time window w_i, we compute the following statistical features along all 25 dimensions for each session i:

- Standard deviation (std)
- Mean value (mean)
- Area under the curve value (auc)

Since the data points of the action units represent intensities, peaks in the respective time series indicate how clear (and how often) the corresponding cue was shown. In order to include this information into the classification process, we additionally compute the following two features related to maximum values for each action unit intensity:

- Maximum value (max)
- Number of peaks higher than the mean value (peaks)

This results in a feature set with 109 features per window. In addition, we also encode the reported stance of the participant on the overall topic into the feature set. As the stance was reported on a scale from 1 (completely disagree) to 5 (fully agree), the numerical rating directly serves as a feature, resulting in a feature vector $\mathbf{x}_i = (x_i^1, \dots, x_i^n)$ with $n = 110$ entries. As we have included features with different ranges, we re-scale each feature as

$$x^k = \frac{x^k - min(x^k)}{max(x^k) - min(x^k)} \quad (1)$$

where x^k denotes the list of feature values along dimension k. A well-known problem of this re-scaling method is that extreme outliers lead to a compression of a majority of the data points into a very small range. However, in the present case this is not an issue as we are using mostly statistical features that already put isolated extreme data points into perspective. The only exception is the maximum value of each action unit, which is provided by OpenFace on a fixed scale from 0 to 5 and therefore can also not include extreme outliers.

VII. EXPERIMENTAL SETUP AND RESULTS

Within this section, we discuss our approach to estimating the participant opinion based on the extracted features introduced in the previous section. We see that we have data sets of 610 (*convincing*) and 653 (*interesting*) sessions as well as 110 features per session. Based on these numbers, we expect the following challenges and requirements regarding the underlying machine learning problem:

- In comparison to other machine learning problems and corpora, the available data sets are small and we

therefore only consider data-efficient methods, namely Support Vector Machine (SVM) and Random Forest (RF).
- The ratio of sessions and features indicates a high probability for overfitting during the training.
- The class distribution for the *interesting* task requires additional pre-processing to avoid majority vote classification.

Throughout this work, we use two different metrics to measure the performance of our model which are the percentage of correct classifications, i.e., the accuracy (ACC) and the unweighted average recall (UAR). Especially in the *interesting* case, where the classes are unequally distributed, the UAR will be the most informative metric for the overall performance.

A. CLASS BALANCE

Especially in case of the *interesting* task, the low number of negative samples is likely to result in a majority class model, i.e., a model that always chooses the majority class regardless of the corresponding feature. In order to prevent that, we utilize two different approaches.

- Class weights adapted to the actual class balance (only SVM)
- Synthetic Minority Oversampling Technique [35] (SVM and Random Forest)

Support Vector Machines separate data points by means of the hyperplane in a large-dimensional space that maximizes the margin between the plane and the data points [36]. The soft-margin parameter C regulates the cost of wrong classifications and represents the trade-off between a large margin and the number of wrongly classified training samples. In the case of balanced data, this value is symmetrical for both classes, meaning that the margin has the same size on both sides of the hyperplane. For imbalanced data, the C value can be adapted to the class balance in order to give the underrepresented samples additional weight [37]. Throughout this work, we use the balanced value $C_{pos,neg} = w_{pos,neg}C$ with

$$w_{pos,neg} = \frac{\#samples}{2 \times \#samples(pos/neg)} \tag{2}$$

where $\#samples$ denotes the number of all samples in the training set, $\#samples(pos)$ the number of samples with a positive rating and $\#samples(neg)$ the number of samples with a negative rating. Therefore, the cost for a wrong classification is increased for samples in the minority class.

The second approach to address the class imbalance is called Synthetic Minority Oversampling Technique (SMOTE). The simplest way of oversampling only replicates data points in the minority class and thereby increases its overall numbers of samples. This approach is extended in SMOTE by generating synthetic data points of the minority class which are close to the original ones in the feature space. In doing so, additional and new data points of the minority class are generated and the overall training set can be balanced. As this approach is applied to the data set, it is independent of the utilized classifier.

TABLE 4. Results of SVM and Random Forest on the *convincing* task for three different feature sets.

Model	Metric	Full	Eye	AU
SVM	ACC UAR	0.61 0.61	0.63 0.63	0.57 0.57
RF	ACC UAR	0.61 0.60	0.61 0.61	0.57 0.56

TABLE 5. Results of SVM and Random Forest on the *interesting* task for three different feature sets. The imbalanced data is addressed by SMOTE oversampling (SMOTE) and an adjusted class weight (cw) in the SVM.

Model	Metric	Full	Eye	AU
SVM + cw	ACC UAR	0.66 0.64	0.65 0.64	0.68 0.61
SVM + SMOTE	ACC UAR	0.67 0.61	0.62 0.63	0.67 0.59
RF + SMOTE	ACC UAR	0.66 0.63	0.64 0.61	0.65 0.61

B. RESULTS

To avoid overfitting as much as possible, we utilize repeated k-fold cross-validation to evaluate the different machine learning models. More precisely, we averaged the results of five 10-fold cross-validations to compute the final validation score for each investigated configuration. The model parameters of both SVM and Random Forest are selected with the complete feature set in a systematic grid search for each task separately. Afterwards, we divide the features into three groups, namely the complete feature set (full, 110 features), a feature set without action units (eye, 25 features) and feature set without eye movements (AUs, 87 features) and investigate the model performance for each feature and task set separately. The results for both models, the *convincing* task and all three feature sets are shown in Table 4.

We see from the results that the SVM performs slightly better than the Random Forest classifier and that the performance of the SVM is actually increased on the limited feature set that includes no information about the AUs. In addition, we see that both metrics are very similar for all the cases, which can be attributed to the almost equal class balance in the data.

For the *interesting* task, we compare both machine learning techniques for all three data sets as well as the approaches discussed in Section VII-A to deal with the unequal class balance. The results can be seen in Table 5.

In terms of UAR, the SVM with the adjusted class weight performs best. Since the UAR accounts for both classes equally and is hence independent of the class balance, we consider it the most relevant metric for this task and it is therefore used as the main indicator of the performance. In the next step, we compare the results of the repeated 10-fold cross-validation to a leave-one-participant-out cross-validation in order to investigate if the cues shown by the participants are individually different. The results for both SVM and Random Forest in the case of the *convincing* task for all three feature sets are shown in Table 6. We see a clear decrease in both metrics for the facial action unit feature set. Moreover, the SVM again outperforms the Random Forest classifier and the best results are similar to the ones achieved in the repeated 10-fold cross-validation.

TABLE 6. Results of SVM and Random Forest on the *convincing* task for three different feature sets and leave-one-participant-out cross-validation.

Model	Metric	Full	Eye	AU
SVM	ACC UAR	0.60 0.60	0.62 0.62	0.51 0.51
RF	ACC UAR	0.61 0.60	0.58 0.58	0.52 0.52

The results for the *interesting* task, all three classifier configurations and all three feature sets are shown in Table 7. Again, we see a clear drop in both metrics for the facial action unit feature set but in contrast to the *convincing* task, also for the full feature set. In this case, the performance of the different classifier configurations varies and there is no clear advantage for one approach over all three feature sets. However, the overall best results (in terms of UAR) are again achieved with the class-weighted SVM.

TABLE 7. Results of SVM and Random Forest on the *interesting* task for three different feature sets and leave-one-participant-out cross-validation. The imbalanced data is addressed by SMOTE oversampling (SMOTE) and an adjusted class weight (cw) in the SVM.

Model	Metric	Full	Eye	AU
SVM + cw	ACC UAR	0.55 0.52	0.63 0.61	0.52 0.46
SVM + SMOTE	ACC UAR	0.57 0.49	0.59 0.58	0.55 0.45
RF + SMOTE	ACC UAR	0.62 0.56	0.60 0.53	0.59 0.53

We conclude that the classification is in some instances hindered or at least not supported by a large set of features, which can be attributed to the limited amount of available data. In addition, the eye movement data appears to be more informative than the facial action units, especially in the *convincing* task. For a comparison with the performance of human annotators in the following subsection, we use the SVM with the feature sets eye (*convincing)* and full (*interesting*) as they perform best in the most general 10-fold cross-validation setup.

VIII. HUMAN ANNOTATION

To compare the machine learning approaches to human performance, we conducted an annotation on a subset of the recordings (10%). Due to the high variations observed during the individual 10-fold cross-validations, this test set was selected in compliance with the following conditions from a list of randomly generated candidate sets to ensure a representative comparison:

- The class balance for both tasks in the test set is representative for the class balance of the overall dataset.
- The deviation of the machine learning performance on the test set from the average performance is lower or equal than the average standard deviation of the 10-fold cross-validation.

Annotators watched snippets of the experiment recordings and were asked to rate if the observed person is interested in and convinced by the presented arguments. In addition, annotators were asked to report their confidence in each rating on a scale from 1 (low confidence) to 5 (high confidence). Annotators had the same information available that are used for the classification, namely the video snippet in the period from the beginning of the system utterance up to the first user rating and the reported stance of the respective participant on the discussed topic. In addition, annotators answered two questionnaires, one before starting the annotation and one after completion. In the first questionnaire, annotators were asked to name the indicators they assume to be the most influential and to rate the sentence *The task will be difficult* on a five-point Likert scale. After completing the task, annotators were interviewed again regarding the most influential indicators for both tasks and their opinion on the difficulties they encountered. Each annotator also answered the question *The task was difficult* again on a five-point Likert scale.

TABLE 8. Performance of the human annotators (A1-A3) and the SVM model on the test set for the *convincing* category.

Metric	A1	A2	A3	Majority	SVM
ACC UAR	0.64 0.63	0.67 0.67	0.64 0.64	0.67 0.67	0.64 0.64

The annotation was done by three annotators (2 male, 1 female) at NAIST for each session in the test set. We measure the inter-annotator agreement for both tasks by means of Fleiss' Kappa [38]. For the *convincing* category, we report an agreement between the human annotators of $\kappa = 0.24$ with a corresponding p-value of $p = 0.001$ which is a fair agreement. The accuracy and UAR values for all three human annotators, the majority rating and the SVM model are shown in Table 8.

In case of the *interesting* category, the inter-annotator agreement between the human annotators is $\kappa = 0.01$ with a p-value of $p = 0.864$, which means that the agreement is not more than random. The results for each annotator, the majority rating and the SVM model for this task are shown in Table 9.

Overall we observe a similar performance of the SVM and the human annotation in both tasks. In order to investigate the differences between the two, we additionally compare

TABLE 9. Performance of the human annotators (A1-A3) and the SVM model on the test set for the *interesting* task.

Metric	A1	A2	A3	Majority	SVM
ACC	0.78	0.64	0.60	0.72	0.67
UAR	0.55	0.44	0.53	0.52	0.60

TABLE 10. Class-wise recall of the human majority rating and the SVM classification for both tasks.

Category	Class	SVM	Majority
Convincing	positive	0.66	0.59
	negative	0.63	0.75
Interesting	positive	0.73	0.89
	negative	0.46	0.15

the class-wise recall of the human majority rating and the SVM classification for both tasks. The corresponding results are shown in Table 10. For the *convincing* task, we see that the SVM has similar performance for both classes, whereas the human ratings are better for negative samples. In case of the *interesting* task, both approaches yield better results for the positive class. In comparison, the SVM outperforms the human annotation in the negative class whereas the majority rating of the human annotators is better than the SVM for the positive class.

As for the questionnaires, all three annotators rated the expected difficulty of the task (before the experiment) with 3 and the actual difficulty (after the experiment) with 4. Indicating cues that were mentioned by at least two annotators were eye movements and facial expressions for the *interesting* task as well as facial expressions and head movements for the *convincing* task. For the *interesting* task, these cues are in line with the features utilized in the SVM, whereas the cues for the *convincing* task differ from the eye movement features used in the machine learning approach.

In addition, two of the annotators reported a lack of reactions in some instances and all three annotators said that they would assume more expressions in a human discussion with a more active role of the user. Moreover, two of the three annotators reported that it was helpful to see more than one clip of a person. Finally, the average confidence scores for both tasks and all three annotators are shown in table 11. It can be seen that despite the lower agreement in the *interesting* case, the average confidence is higher than in the *convincing* case for the annotators A1 and A2.

TABLE 11. Average confidence of the human annotators (A1-A3) for both categories.

Category	A1	A2	A3
Convincing	3.23	2.56	3.74
Interesting	3.67	2.78	3.33

IX. DISCUSSION

Throughout this section, we discuss our findings and possible implications. In general, we can see that all results are clearly above the random guess baseline of 50% and hence, that the tasks can be addressed by machine learning approaches. However, the low amount of (imbalanced) data and the observed high variability in the classification results makes further general assertions difficult. We start the detailed discussion with the comparison of human to machine learning performance, followed by a comparison of our results to literature values. Finally, we look at the results from the perspective of applications in future argumentative dialogue systems.

A. COMPARISON OF ANNOTATION AND CLASSIFICATION

Overall, we observe similar results for the human annotation and the best machine learning approach (SVM) in both tasks. This indicates that the reported performance is close to the upper bound of the utilized data, which is also in line with the comments of the annotators that reported a lack of cues in several instances. However, the comparison of the class-wise recall shows that there are also differences between the human annotation and the machine classification, i.e. that some instances are classified correctly by one approach and not by the other. This can in parts be attributed to the technical configuration of the SVM, especially in the *interesting* task where the imbalanced data and the adjusted class weight of the SVM lead to a better performance in detecting negative samples at the cost of lower performance in the detection of positive ones. Also, the indicating cues reported most frequently by the annotators for this task are in line with the features used in the SVM classification, which indicates that similar information is used by both approaches. For the *convincing* task however, the observed differences in combination with the reported indicating cues suggest, that different information is used by the two compared approaches and hence, that additional or alternative features might improve the machine learning performance further. If and to what extent such an improvement is possible will be investigated on a larger data set in future work. As also reported by the annotators and indicated by related work, a more active role of the user is likely to improve the results as well since it allows the use of additional information (like gestures and linguistic features) and leads to a more expressive behaviour of the participants in general. In addition, the results also indicate that a user can hide his or her opinion from the system recognition.

Regarding the two different tasks, the results of the human annotation show that the *interesting* category appears to be more challenging. One reason for the difficulty of this task can be attributed to the imbalanced class distribution and the lack of clearly negative examples. Moreover, the experimental setup is also likely to influence the results, as one annotator (A3) reported the impression that participants try to be generally interested as a consequence of the recording (and not necessarily the arguments). The presence of cues that are related to general (dis)interest and not to the arguments is also an explanation for the higher average confidence scores of A1 and A2 in this task, which are in contrast to the low agreement between the annotators. This is an important factor for applications as well since the reason for the shown response cannot always be identified. From this perspective, it would also be beneficial to have a more active role of the user, as well as confirmation strategies on the system side that can help to clarify the situation.

B. COMPARISON WITH LITERATURE

As discussed in Section II, similar approaches in other domains/applications were investigated and we compare our results to the setups that are most similar to ours. For the *convincing* task, we compare our results to the assessment of spontaneous (dis)agreement by means of nonverbal cues [20]. Similar to our case, the authors investigate spontaneous (dis)agreement recognition and address it as a binary class problem. In contrast to the herein presented approach, the data set is based on human-human debates in which the observed individuals assume an active role and the data label is based on human annotation. In addition, prosodic features i.e. features extracted from the voice of the speakers and visual features related to hand actions as well as head and body gestures are utilized for the estimation. The reported overall accuracy of 64% is close to the herein reported accuracy of 63%, although it was accomplished with both prosodic and visual features. The reported accuracy without prosodic features is only slightly above 50% (exact values are only provided for the best results) and hence clearly lower than the herein achieved 63%.

In the case of interest recognition, we compare our results to the ones presented in [8]. The similarity to our task is mainly due to the very similar setup of human-machine interaction as well as a turn-wise assessment of the user interest on a binary scale. The differences to our work are the investigated domain, the annotation-based labelling of the data and the use of both prosodic and visual features for classification. The best reported average recall score of 70% is higher than the herein reported 64%. Nevertheless, the authors also report an average recall of 62% achieved without prosodic features which indicates that the herein achieved results are comparable to the ones in the referenced scenario.

We conclude that the comparison with the literature yields similar or better results than achieved in the reference work without prosodic features. However, both comparisons also indicate that the herein reported performance could be improved by using prosodic features in a scenario with a more active user where those are accessible. In addition, a more active role of the user would also allow for an exploration of additional visual features such as gestures and postures for the herein considered tasks.

C. APPLICATIONS AND LIMITATIONS

For applications, it is evident that the herein discussed approaches need to be enhanced to perform reliably. In particular, additional training data is required to ensure a robust recognition in application scenarios. Also, users are able to hide their opinion which can hinder the recognition in competitive tasks like negotiation even further. Consequently, scenarios in which the user cooperates with the system are suitable candidates for first applications. For example, a system like the one introduced in [4] for the information-seeking domain could use the herein proposed methods in combination with confirmation strategies in direct interaction with human users. In addition, the more active role of the user in such a scenario is likely to facilitate the recognition of both herein discussed quality aspects.

A remaining open question is whether or not a personalized model of the user opinion established over multiple interactions can further improve the recognition. A comparison of the best results for the 10-fold cross-validation and the leave-one participant out cross-validation shows only a very small difference. However, these results correspond to a classification solely based on eye movement features, whereas especially for the facial action units, the results clearly decrease in the leave-one-participant-out cross-validation. A possible explanation is that the facial expressions are more individual and the eye movements are more general. This is also in line with the somewhat controversial reports of the annotators: One annotator said that additional clips of a single participant did not help in accomplishing the task as no context information is provided thereby. In contrast, the remaining two annotators reported that additional clips of the participants helped.

X. CONCLUSION

We discussed the estimation of subjective argument quality criteria from social signals, namely how *convincing* and how *interesting* an argument was perceived by a human user. In order to do so, we utilized data collected for the evaluation of argument search engines with a dialogue system that allows users to rate single arguments in multiple categories. The corresponding ratings were estimated from the facial expressions and the eye movements shown by the participants during the interaction with the system.

A comparison with human annotations yielded similar results, with a slight advantage of the human annotation in the *convincing* task and a slight advantage of the machine learning approach in the *interesting* task (in terms of UAR). Moreover, we reported a fair agreement between human annotators for the *convincing* case, whereas the *interesting* case showed only a random agreement. Finally, we compared our results to the ones achieved in the literature for scenarios that bear similarities to the herein addressed tasks. The comparison showed that our results are similar (*interesting*) or above (*convincing*) the results reported in the compared literature for similar feature sets (visual features). In summary, we draw the following conclusions:

- The overall results are clearly above the random baseline and comparable to performances reported in the literature for similar tasks and feature sets. Hence, a recognition of both subjective argument quality aspects from social signals is feasible.

- The results of the human annotation show that both tasks are nevertheless challenging and additional data in combination with a more active role of the user is likely to improve the performance.
- The machine learning results are in the same range as the performance of the human annotators and it is therefore unlikely that the results can be significantly improved by additional features in the current setup.

Since one limitation of the present work is the limited amount of training data, the first step in future work will be an extension of the discussed experiment with more participants. Moreover, we will investigate if the machine learning performance can be improved if the user assumes a more active role in the interaction and include additional features that are accessible in such a case. In addition, we will look at techniques that can utilize the estimated information for adaptation, as for example Reinforcement Learning [39].

REFERENCES

[1] L. A. Chalaguine, A. Hunter, H. Potts, and F. Hamilton, "Impact of argument type and concerns in argumentation with a chatbot," in *Proc. IEEE 31st Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2019, pp. 1557–1562.

[2] A. Rosenfeld and S. Kraus, "Strategical argumentative agent for human persuasion," in *Proc. 22nd Eur. Conf. Artif. Intell.* Amsterdam, The Netherlands: IOS Press, 2016, pp. 320–328.

[3] B. Galitsky, "Enabling a bot with understanding argumentation and providing arguments," in *Developing Enterprise Chatbots*. Cham, Switzerland: Springer, 2019, pp. 465–532.

[4] A. Aicher, N. Rach, W. Minker, and S. Ultes, "Opinion building based on the argumentative dialogue system Bea," in *Proc. 10th Int. Workshop Spok. Dialog Syst. Technol. (IWSDS)*, 2019.

[5] H. Wachsmuth, N. Naderi, Y. Hou, Y. Bilu, V. Prabhakaran, T. A. Thijm, G. Hirst, and B. Stein, "Computational argumentation quality assessment in natural language," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, vol. 1, 2017, pp. 176–187.

[6] N. Rach, Y. Matsuda, J. Daxenberger, S. Ultes, K. Yasumoto, and W. Minker, "Evaluation of argument search approaches in the context of argumentative dialogue systems," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 513–522.

[7] K. Weber, K. Janowski, N. Rach, K. Weitz, W. Minker, S. Ultes, and E. Andrè, "Predicting persuasive effectiveness for multimodal behavior adaptation using bipolar weighted argument graphs," in *Proc. 19th Int. Conf. Auton. Agent Multi-Agent Syst.*, 2020, pp. 1476–1484.

[8] S. Tomimasu and M. Araki, "Assessment of users' interests in multimodal dialog based on exchange unit," in *Proc. Workshop Multimodal Analyses Enabling Artif. Agents Hum.-Mach. Interact.* New York, NY, USA: ACM, Nov. 2016, pp. 33–37.

[9] K. Bousmalis, M. Mehu, and M. Pantic, "Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools," *Image Vis. Comput.*, vol. 31, no. 2, pp. 203–221, Feb. 2013.

[10] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of facial expressions and measurement of levels of interest from video," *IEEE Trans. Multimedia*, vol. 8, no. 3, pp. 500–508, Jun. 2006.

[11] A. Lauscher, L. Ng, C. Napoles, and J. Tetreault, "Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing," 2020, *arXiv:2006.00843*. [Online]. Available: http://arxiv.org/abs/2006.00843

[12] I. Habernal and I. Gurevych, "What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in Web argumentation," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Austin, TX, USA: ACL, 2016, pp. 1214–1223.

[13] M. Gleize, E. Shnarch, L. Choshen, L. Dankin, G. Moshkowich, R. Aharonov, and N. Slonim, "Are you convinced? Choosing the more convincing evidence with a siamese network," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 967–976.

[14] H. Wachsmuth, N. Naderi, I. Habernal, Y. Hou, G. Hirst, I. Gurevych, and B. Stein, "Argumentation quality assessment: Theory vs. Practice," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2. Vancouver, BC, Canada: ACL, 2017, pp. 250–255.

[15] A. Toledo, S. Gretz, E. Cohen-Karlik, R. Friedman, E. Venezian, D. Lahav, M. Jacovi, R. Aharonov, and N. Slonim, "Automatic argument quality assessment–new datasets and methods," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 5624–5634.

[16] S. Gretz, R. Friedman, E. Cohen-Karlik, A. Toledo, D. Lahav, R. Aharonov, and N. Slonim, "A large-scale dataset for argument quality ranking: Construction and analysis," in *Proc. 34th AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 7805–7813.

[17] M. Potthast, L. Gienapp, F. Euchner, N. Heilenkötter, N. Weidmann, H. Wachsmuth, B. Stein, and M. Hagen, "Argument search: Assessing argument relevance," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 1117–1120.

[18] I. Poggi, F. D'Errico, and L. Vincze, "Agreement and its multimodal communication in debates: A qualitative analysis," *Cognit. Comput.*, vol. 3, no. 3, pp. 466–479, Sep. 2011.

[19] K. Bousmalis, M. Mehu, and M. Pantic, "Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools," in *Proc. 3rd Int. Conf. Affect. Comput. Intell. Interact. Workshops*, Sep. 2009, pp. 1–9.

[20] K. Bousmalis, L.-P. Morency, and M. Pantic, "Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition," in *Proc. Face Gesture*, Mar. 2011, pp. 746–752.

[21] B. Schuller, R. Müeller, B. Höernler, A. Höethker, H. Konosu, and G. Rigoll, "Audiovisual recognition of spontaneous interest within conversations," in *Proc. 9th Int. Conf. Multimodal Interfaces (ICMI)*. New York, NY, USA: ACM, 2007, pp. 30–37, doi: 10.1145/1322192.1322201.

[22] J. H. Jeon, R. Xia, and Y. Liu, "Level of interest sensing in spoken dialog using decision-level fusion of acoustic and lexical evidence," *Comput. Speech Lang.*, vol. 28, no. 2, pp. 420–433, Mar. 2014.

[23] Y. Sasaka, T. Ogawa, and M. Haseyama, "Multimodal interest level estimation via variational Bayesian mixture of robust CCA," in *Proc. 24th ACM Int. Conf. Multimedia*. New York, NY, USA: ACM, Oct. 2016, pp. 387–391.

[24] T. Hirayama, J.-B. Dodane, H. Kawashima, and T. Matsuyama, "Estimates of user interest using timing structures between proactive content-display updates and eye movements," *IEICE Trans. Inf. Syst.*, vol. 93, no. 6, pp. 1470–1478, 2010.

[25] T. Ahmed, R. Singh, A. K. Pandey, and S. K. Singh, "A cognitive model to predict human interest in smart environments," *Comput. Commun.*, vol. 161, pp. 1–9, Sep. 2020.

[26] Y. Ajjour, H. Wachsmuth, J. Kiesel, M. Potthast, M. Hagen, and B. Stein, "Data acquisition for argument search: The args. me corpus," in *Proc. Joint. German/Austrian Conf. Artif. Intell. (Künstliche Intell.)*. Cham, Switzerland: Springer, 2019, pp. 48–59.

[27] J. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre, and M. Cieliebak, "Survey on evaluation methods for dialogue systems," *Artif. Intell. Rev.*, pp. 1–56, Jun. 2020.

[28] C. Reed and T. Norman, Eds., *Argumentation Machines: New Frontiers in Argument and Computation*, vol. 9. Dordrecht, The Netherlands: Kluwer, 2003.

[29] H. Ritschel, T. Baur, and E. André, "Adapting a robot's linguistic style based on socially-aware reinforcement learning," in *Proc. 26th IEEE Int. Symp. Robot Hum. Interact. Commun. (RO-MAN)*, Aug. 2017, pp. 378–384.

[30] C. Stab, J. Daxenberger, C. Stahlhut, T. Miller, B. Schiller, C. Tauchmann, S. Eger, and I. Gurevych, "ArgumenText: Searching for arguments in heterogeneous sources," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Demonstrations*, 2018, pp. 21–25. [Online]. Available: http://tubiblio.ulb.tu-darmstadt.de/105466/

[31] H. Wachsmuth, M. Potthast, K. A. Khatib, Y. Ajjour, J. Puschmann, J. Qu, J. Dorsch, V. Morari, J. Bevendorff, and B. Stein, "Building an argument search engine for the Web," in *Proc. 4th Workshop Argument Mining*, 2017, pp. 49–59.

[32] N. Rach, K. Weber, L. Pragst, E. André, W. Minker, and S. Ultes, "EVA: A multimodal argumentative dialogue system," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 551–552.

[33] A. F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," *Commun. Methods Measures*, vol. 1, no. 1, pp. 77–89, Apr. 2007.

[34] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 59–66.

[35] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[36] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 2000.

[37] X. Yang, Q. Song, and Y. Wang, "A weighted support vector machine for data classification," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 21, no. 5, pp. 961–976, Aug. 2007.

[38] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, p. 378, 1971.

[39] K. Weber, N. Rach, W. Minker, and E. André, "How to win arguments," *Datenbank-Spektrum*, vol. 20, no. 2, pp. 161–169, Jul. 2020.

NIKLAS RACH studied physics at Ulm University, Germany. He received the M.Sc. degree in 2015. He is currently pursuing the joint Ph.D. degree in computer science with the Dialogue Systems Group, Ulm University, and the Ubiquitous Computing Systems Laboratory, Nara Institute of Science and Technology, Japan. Subsequently, he spent three months as a Visiting Student at the Center of Theoretical Atomic, Nuclear and Optical Physics, Queen's University, Belfast, U.K., from November 2015 to January 2016. His research interests include dialogue management and decision making with an emphasis on computational argumentation in dialogue systems and machine learning applications.

YUKI MATSUDA (Member, IEEE) was born in 1993. He received the B.E. degree from the Advanced Course of Mechanical and Electronic System Engineering, National Institute of Technology, Akashi College, Japan, in 2015, and the M.E. and Ph.D. degrees from the Graduate School of Information Science, Nara Institute of Science and Technology, Japan, in 2016 and 2019, respectively. Since 2019, he has been an Assistant Professor of the Ubiquitous Computing Systems Laboratory, Graduate School of Science and Technology, Nara Institute of Science and Technology. Since 2020, he has also been a Researcher of the Japan Science and Technology Agency PRESTO. His current research interests include participatory sensing, location-based information systems, wearable computing, and affective computing. He is currently a member of IPSJ.

STEFAN ULTES received the diploma (M.Sc.) degree in computer science from the Karlsruhe Institute of Technology, Germany, in 2010, and the Ph.D. degree in engineering from the Dialogue Systems Group, Ulm University, Germany, in 2015. Afterwards, he was a Research Associate with the Spoken Dialogue Systems Group, University of Cambridge, working with Prof. S. Young and Prof. M. Gasic within the EPSRC Project "Open Domain Statistical Spoken Dialogue Systems." He is currently employed as a Dialogue Research Lead with Mercedes Benz Research and Development working on the next generation of the Mercedes Benz User Experience.

WOLFGANG MINKER received the diploma (M.Sc.) and Ph.D. degrees in engineering science from the University of Karlsruhe, Germany, in 1997, and the Ph.D. degree in computer science from Université Paris-Sud, France, in 1998.

From 1993 until 2000, he was a Teaching Assistant with LIMSI-CNRS, Université Paris-Sud, France, and subsequently worked as a Senior Researcher with the Dialogue Systems Group of DaimlerChrysler Research and Technology, Ulm, Germany, from 2000 to 2003. Since 2003, he has been a Full Professor with Ulm University, Germany, and also became a Co-Director of the International Research Laboratory "Multimodal Biometric and Speech Systems," ITMO University St. Petersburg, Russia, in 2017. The research at his group is focused on dialogue systems with special interests in spoken dialogue interaction in ambient intelligent environments, assistive, adaptive, and proactive spoken language dialogue interaction, dialogue modeling, and argumentative dialogue systems.

KEIICHI YASUMOTO (Member, IEEE) received the B.E., M.E., and Ph.D. degrees in information and computer sciences from Osaka University, Osaka, Japan, in 1991, 1993, and 1996, respectively. He is currently a Professor with the Graduate School of Science and Technology, Nara Institute of Science and Technology. His research interests include distributed systems, mobile computing, and ubiquitous computing. He is also a member of ACM, IPSJ, SICE, and IEICE.