Let's move on: Topic Change in Robot-Facilitated Group Discussions

Georgios Hadjiantonis[1], Sarah Gillet[1], Marynel Vázquez[2], Iolanda Leite[1] and Fethiye Irmak Dogan[1]

***Abstract*— Robot-moderated group discussions have the potential to facilitate engaging and productive interactions among human participants. Previous work on topic management in conversational agents has predominantly focused on human engagement and topic personalization, with the agent having an active role in the discussion. Also, studies have shown the usefulness of including robots in groups, yet further exploration is still needed for robots to learn when to change the topic while facilitating discussions. Accordingly, our work investigates the suitability of machine-learning models and audiovisual non-verbal features in predicting appropriate topic changes. We utilized interactions between a robot moderator and human participants, which we annotated and used for extracting acoustic and body language-related features. We provide a detailed analysis of the performance of machine learning approaches using sequential and non-sequential data with different sets of features. The results indicate promising performance in classifying inappropriate topic changes, outperforming rule-based approaches. Additionally, acoustic features exhibited comparable performance and robustness compared to the complete set of multimodal features. Our annotated data is publicly available at https://github.com/ghadj/topic-change-robot-discussions-data-2024.**

I. INTRODUCTION

Group discussions are commonly used for brainstorming and making informed decisions, promoting collaboration, diversity, and innovative thinking [1]. However, managing the conversation flow can be challenging for the participants. In such cases, a moderator can play an essential role in ensuring that the discussion is engaging, the participants establish a common ground, and transitions between discussion points are smooth and timely. In this work, we explore how a robot could facilitate a group discussion by deciding *when* a transition between topics is needed and appropriate.

Recent research in group Human-Robot Interaction (HRI) has highlighted the impact of robot behavior on the quality and effectiveness of the interaction, as well as the ability of the robot to influence the individuals and shape group dynamics [2], [3]. Such examples demonstrate that robots can have a positive impact on the level of verbal communication among adults in care facilities [4], [5], balance participation during team decision-making discussions [6], and improve task performance and group cohesion [7], [8].

This work investigates how robot facilitators can autonomously decide when to change a discussion topic. Unlike prior work, which typically decides when to change topic through a Wizard-of-Oz paradigm [9], [10] or by using handcrafted heuristics [11], our work focuses on endowing robots with the ability to moderate the topic of a discussion between humans. Using a group HRI dataset collected with the setup depicted in Figure 1, we explore the possibility of having robots decide in a fully autonomous manner on topic changes during discussions, as long as the robot does not speak over people.

Fig. 1. The interaction between the robot and three participants from two different perspectives. The robot moderates the group discussion and needs to decide when to move on to the next topic.

We frame the problem of deciding on topic changes as a classification task, i.e., the robot decides, given a set of multimodal features, if a topic change is *needed*, *appropriate*, or *inappropriate*. We investigate leveraging information beyond verbal cues in this work and evaluate a variety of models in a content-free manner that is not limited to transcription or speech recognition of the conversations. In particular, our set of features is informed by research concerning content-free approaches to topic segmentation [12], [13], [14] and related linguistic research on cues that describe the structure of discourse topics and turn-taking. We evaluate these multimodal features on their promise for this problem and study varying machine learning models using a time sequence of feature data (sequential modeling) and aggregated individual feature data (non-sequential modeling) for topic change prediction.

To our knowledge, this paper is the first to consider the problem of decision-making for changes in a discussion topic in robot-moderated discussions. In summary, our work makes the following contributions:

– We address an unsolved decision-making problem in HRI: when should a robot moderator change the topic of a group discussion among multiple people?

[1]KTH Royal Institute of Technology, Sweden, Contact: {ghad, sgillet, iolanda, fidogan}@kth.se. This work was supported by the S-FACTOR project from NordForsk, the Swedish Foundation for Strategic Research (SSF FFL18-0199), and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

[2]Yale University, USA, Contact: marynel.vazquez@yale.edu. Marynel was supported by the National Science Foundation (IIS-2143109).

– We evaluate multi-modal features and assess their importance to the multiparty topic-change problem.
– We investigate how to apply sequential and non-sequential input-based machine learning models to the topic-change problem and evaluate their performance.
– We contribute an annotated dataset for benchmarking topic-change prediction algorithms and expanding research on this problem domain.

II. Related work

The study of robots and groups has gained importance in HRI [15], including how robots shape and facilitate group dynamics [2], [3]. In particular, robots have been shown to improve conflict situations [16], [17], provide emotional support [18], and foster the expression of vulnerability [19] and group cohesion [7]. Further, prior work studied how robots could support the process of inclusion among adults [20], and children [21], [22], shape participation behavior [6], [23], moderate collaborative games [24], [8], facilitate educational activities [11] or brainstorming sessions [10], [25].

In group settings, discussion topics play a critical role in people's involvement in the interactions. The problem of selecting a discussion topic has traditionally been studied with conversational agents which aim to cohesively select appropriate topics [26] or personalize them based on user engagement [27]. Other work has focused on how to guide conversations into a target subject [28] or to bridge different topics [29]. While our work addresses aspects of conversational management, it does not involve topic selection. Our work focuses on predicting *when* a topic should change.

Topic change is a critical aspect of ensuring people's engagement in group discussions, which can be derived from verbal cues and non-verbal indicators [30], [31], [32]. Recent studies on topic segmentation have suggested multi-modal approaches and acoustic-based methods, using prosodic and visual features [12], [13], [14], [33]. In parallel, linguistic research has suggested a variety of cues indicating topic boundaries and turn-taking, including prosodic features [34], [35], [36], speech rate [37], body gestures [38], [39], [40], [41], [42], and gaze [43], [44], [45]. Our work builds upon the prior understanding of the features that are correlated with the end of discussion topics for determining appropriate topic changes. Additionally, unlike topic segmentation, our approach does not utilize a fully completed conversation but only uses momentary information from the discussion to proactively make real-time topic change decisions.

III. Dataset

We used a dataset collected from interactions between the Shutter robot [46] and groups of human participants (16 groups of two, and nine groups of three participants) as depicted in Figure 1. The group was asked to brainstorm about robots in home environments, while the robot had the role of moderating the discussion. Data was collected through individual close-talk microphones and the body tracking module of an Azure Kinect Camera, which was placed behind the robot.

During the discussions, the robot presented various topics and asked leading questions to guide the discussion. For example, the robot would ask: *"What do you think about robots like me being around in home environments?"* Further, the robot could ask follow-up questions to encourage more ideas and deepen the discussion, e.g., *"Do you have other ideas to share?"*. To determine whether to move on to the next topic or ask a follow-up question, the robot used a simple rule-based heuristic devised by an expert: a topic change would be initiated after a total of 60 seconds of speech when none of the participants was speaking (not speaking was based on a 2 seconds silence threshold). We use this heuristic as a baseline method in this work (described in Section IV-D).

A. Annotations

The dataset was manually annotated by the first author in order to identify whether, at the end of each utterance, the robot should change the topic (i.e., the change is ***needed***), could change the topic (the change is ***appropriate***) or should wait for more contributions (the change is ***inappropriate***). The annotator watched the video recordings from the robot's point of view and focused on labeling robot decisions two seconds after an utterance ended. This approach ensured that the robot would wait until human participants concluded talking and allowed it to capture valuable insights from the moments directly after one participant's contribution.

We used voice activity detection with a silence threshold of 750 msec to detect utterances, which is slightly longer than what was previously used in turn-taking prediction (200 and 500 msec [47]), due to the less demanding response and to account for "search" or "repair pauses", i.e., while a speaker pauses to search for an appropriate word or phrase, or attempt to revise what was previously stated [34]. In total, 1529 utterances were extracted from 2-participant sessions and 930 utterances from 3-participant sessions.

Note that the annotation decision was made considering the whole group interaction, not just considering only the active participant. Figure 2 provides an extract of the interaction between two participants and the robot as well as the corresponding annotation per utterance.

B. Feature Extraction

We created a feature vector for topic-change classification using the data collected from group interactions. The feature vector corresponded to each utterance and included acoustic attributes of the current speaker, hand gestures, body and head movements of all the participants, and the total duration of each utterance. Previous work on turn-taking prediction considered windows of 200-1000 ms at the end of speech for computing acoustic features [48], [49]. Inspired by work in topic change detection, indicating improved performance using an extended context of 2.56 seconds before and after the decision time [12], we experimentally determined to extract all the features during the time interval from 2 seconds before to 2 seconds after the end of each utterance.

We created a fixed-size feature representation for topic-change classification independent of the group size in the

			annotation
1	R:	is there anything that someone could find concerning?	-
2	P2:	concerning? hmm...	not appropriate
3	P1:	about the room or about what?	not appropriate
4	P2:	I think about the robot being in the room	not appropriate
5	P1:	I guess in this type of room there is not a lot of privacy	not appropriate
6	P2:	yeah, the living room is the bedroom, so that is kind of tricky	appropriate
7	P1:	yeah... [both participants turn towards the robot]	appropriate
8	R:	do you have other ideas to share?	-
9	P2:	hmm...	not appropriate
10	P1:	hmm...	not appropriate
11	P2:	hmm... no...	not appropriate
12	P1:	no... I imagine this type of robot in big houses but not really in student rooms or student apartments. uhmm...	not appropriate
13	P1:	I see the limitation there, probably. [both participants turn towards the robot]	needed

Fig. 2. Utterance sequence from our annotated dataset. The interaction is between the robot (R) and two participants (P1 and P2), discussing the question: "Is there any way that a robot like me could be helpful in those places in your home you just described?".

interaction because the dataset contains groups of two and three participants. For acoustic features, we only used the features of the active speaker. For all other features, we used one set of features of the active speaker and a second set as an average over the features of the remaining participants. The features are further detailed below:

Acoustic Features: These features were extracted from the individual audio signals of the active participants. Specifically, we computed the mean, maximum, and minimum value and standard deviation of the speech energy and pitch over the given data window. Additionally, we calculated the mean value of the voice quality features, i.e., jitter, shimmer, and Harmonics-to-Noise Ratio (HNR). Since pitch and voice quality features are characteristics of voice, only the last 2 seconds before the end of the utterance were considered.

Hand Gestures and Body Features: The features were computed from the Kinect body joints. For upper body movements, we calculated the relative position of the chest to the pelvis in the x and y axes, capturing if a participant was leaning forward or sideways, respectively. To capture the hand movements, we calculated the 3D position of the left and right hands, respectively, in relation to the chest. In addition, we included hand velocity and upper body movement by calculating the temporal difference between successive data points. For a detailed definition of the Kinect coordinate system and body tracking joints, refer to the Microsoft Azure Kinect documentation [1].

Head Rotation Features: As a proxy for gaze direction, we computed the relative head rotation between the participants and the robot. First, we measured the relative horizontal head rotation angle between the participants and the robot. To capture the head direction in relation to the interactants independently of their position in the room, the rotation angle between a human and the robot was transformed into one of three values as follows. First, for the human speaker, 0 corresponded to looking at the robot, 1 looking at a listener, and -1 looking away from both the listeners and robot. For a human listener, 0 corresponded to looking at the robot, 1 indicated looking at the speaker, and -1 was looking away in another direction. Similar to the hand and body features, we further calculate the temporal difference of successive head direction values to capture the head rotation movement in addition to head direction.

[1]https://learn.microsoft.com/en-us/azure/kinect-dk/body-joints

IV. METHODOLOGY

A. Problem Formulation

We define the problem of determining when to change the discussion topic at the end of an utterance as a classification problem. The goal is to learn a classification function f that maps input features x to a predicted class label $\hat{y}$:

$$f : x \mapsto \hat{y} \tag{1}$$

where the label $\hat{y} \in \mathcal{Y}$ can take one of three values, $\mathcal{Y} = \{$*not appropriate, appropriate, needed*$\}$. To model the input x, we propose non-sequential and sequential data modeling approaches and explore different functions f handling these data inputs. The two types of approaches are detailed next.

B. Non-sequential Data Modeling Approach

For the non-sequential data modeling approach, each feature is aggregated separately for two-time windows: (1) from 2 seconds before until the end of each utterance, and (2) starting from the end of each utterance until 2 seconds afterwards. This results in an input feature vector $x_i \in \mathbb{R}^N$ for an utterance i, where N is the total number of features.

Using the above features, we approximate the function f from eq. (1) with the following models: a Decision Tree (DT), a Random Forest (RF), a Support Vector Machine (SVM), and a Multilayer Perceptron (MLP) classifier.

The models were trained using the same dataset format, i.e., for M total utterances, $X = [x_1, ..., x_M]$ and $y = [y_1, ..., y_M]$, where $x_i \in \mathbb{R}^N$ and $y_i \in \mathcal{Y}$. Given a new input vector x, the models output a predicted class label $\hat{y} \in \mathcal{Y}$.

Regarding DTs, we used the Gini impurity measure as a split criterion and tuned the max depth of the tree and the min samples per split. Similarly, for RFs, we used the Gini impurity as the split criterion and tuned the number of estimators and maximum depth of trees. For SVMs, we used the Radial Basis Function (RBF) as the kernel function. The kernel function, gamma, and regularization term were tuned (see Section V-D for training and tuning procedure).

For MLPs, we experimented with one and two hidden layers of different sizes with Rectified Linear Unit (ReLU) activation(s) [50]. For the output layer of the network, we used a softmax activation function, estimating the probability distribution over the desired classes. For model supervision, we used the categorical cross-entropy loss. Model parameters were optimized using the Adaptive moment estimation algorithm (Adam) [51]. The batch size used during training was also tuned, considering batch sizes of 8, 16, and 32.

C. Sequential Data Modeling Approach

In contrast to the non-sequential modeling, sequential data is processed as a series of vectors with temporal structure. The sequential features, as the non-sequential ones, were the same types of features and contained information for the period of 2 seconds before until 2 seconds after the end of the utterances. However, instead of aggregating the features on two windows before and after, the sequential features were sampled using a sliding window at a rate of 4 Hz. This resulted in a sequence of input vectors for each utterance, containing the corresponding feature values at each instance, i.e., $x = [x_1, x_2, ..., x_\tau]$, where τ is the length of the sequence between 2 seconds before and after the end of an utterance, and $x_i \in \mathbb{R}^N$, with N the total number of features at each instance.

We used two recurrent models to process the sequential data: Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks. The LSTM and GRU models had one recurrent layer. As a recurrent activation function, we used the sigmoid function, and as activation for the hidden state, the hyperbolic tangent activation function. At the final time step of the recurrent models, the hidden state was fed to a dense layer with a softmax activation, resulting in a probability distribution over the possible classes. The models were trained with a categorical cross-entropy loss. To prevent overfitting, we applied a dropout rate of 0.1 in the input and recurrent connections within the neural networks. The dimensionality of the internal state and the batch size were tuned (from values 8, 16, and 32), and the models were trained with the Adam optimizer.

D. Baselines

1) Feature-Based Heuristic: As a baseline method, we used a set of simple rule-based heuristics which use two threshold values each to predict the three classes. We chose to use a set of heuristics to provide a more general insight into the promise of feature-based heuristics. To choose the features for the heuristics, we performed a forward greedy feature selection approach on the aggregated features with an SVM with RBF kernel [52]. We chose to consider the three top features[2] for the feature-based heuristics given the performance improve during the feature selection process.

We computed the thresholds for each feature-based heuristic separately, maximizing the F1 score on the training set. The final score for the feature-based baseline, as shown in the comparison tables, was calculated as the mean and standard deviation over the performance of the three feature-based heuristics on the test set and holdout set.

2) Speech-and-Pause-Based (SPB) Heuristic: As an additional comparison, we used a similar method to that employed by the robot during the data collection. In contrast with the three classes from the annotation process, the robot's method was used to decide only whether to change the topic. Consequently, for comparison, the method was used to classify the following classes: "not appropriate" and the combination of the other two classes, denoted as "appropriate/needed".

Specifically, an utterance was classified as "appropriate/needed" if the total duration of speech in the current topic was at least 60 seconds, and was followed by a silent pause of at least 2 seconds. Otherwise, the utterance was classified as "not appropriate". Both thresholds were experimentally determined by an expert.

[2]For all sessions combined, the top three features were: (1) maximum speech energy, (2) speaker's head direction, and (3) standard deviation of speech energy

V. Experiments

A. Evaluation Procedure

Our evaluation procedure aimed to answer the following questions:

Q1. *Can machine learning models be used to predict when the topic of the discussion should change? Can they generalize to unseen groups?* We investigated this question in the context of **multi-class classification**, where the models for the sequential data, the models for the non-sequential data and the feature-based heuristic made predictions over three classes ("not appropriate", "appropriate", and "needed"). To better understand the effect of group size on model performance, we analyzed results based on whether the models were trained using all the features for the sessions with two participants, three participants, or all sessions together. Additionally, to investigate the generalization performance of the models, we used two different test sets: one set evaluating the performance of the same groups that are used for training data (test set) and one set with groups that were completely hold-out from training (holdout set).

Q2. *Does two-step classification improve model performance compared to multi-class classification?* We investigated this question by evaluating model performance on **two-step binary classification**. We transformed the problem of 3-class classification (as in Q1) into two binary classification problems: (1) classify "not appropriate" vs. the combined class "appropriate/needed"; and (2) classify "appropriate" vs. "needed". The former binary classification problem was the same setup used by speech-and-pause-based heuristic during the data collection process, as explained in Section IV-D. The latter binary classification problem focused on a more subtle class distinction in order to evaluate if the proposed models could predict the urgency of topic changes.

Q3. *Which features help a robot decide when the discussion should change?* To better understand the value of different features, we evaluated **classification performance using sets of features**: (1) acoustic features, (2) Kinect-derived features, (3) the 20 most important features according to feature selection, and (4) all the available features. To obtain the top 20 features, the greedy feature selection procedure explained for Feature-Based Heuristic in Section IV-D was applied for 20 features. We chose the top 20 features since there were no notable performance improvements after that.

B. Data Splitting and Balancing

The data was partitioned into training, test, and holdout sets. First, a complete session was randomly selected as a holdout set and was balanced by randomly undersampling the majority class. The test set consisted of randomly sampling utterances from the rest of the sessions (each class was determined by the 20% of the size of the minority class). For the training set, the remaining utterances were balanced by selective oversampling of the minority class and randomly undersampling the majority class such that there were an equal number of examples per class. More specifically, the implementation of selective oversampling used the most important feature identified by the forward greedy feature selection (as explained for the Feature-Based Heuristic in Section IV-D): only samples whose selected feature values were between the 25th and 75th percentile were considered during oversampling. Selective oversampling was favored compared to random oversampling to minimize the influence of outliers during training. It is important to note that the test set contains the unseen utterances from groups also used during training; on the other hand, the holdout session quantifies the generalization capabilities to unseen groups.

C. Feature Standardization and Normalization

In order to avoid individual differences among participants, all the features were first standardized using Z-score normalization for each participant. Then, Min-Max normalization was applied so that all the data was scaled in the range between -1 and 1. Normalization parameters were calculated from the training set only to prevent any data leaking.

D. Training and Hyper-parameter Optimization

The hyper-parameters of each model were tuned using grid search. This involved performing 5-fold cross-validation, resulting in five models for each set of parameters. The combination of parameters that achieved the highest mean F1-score on the validation folds was selected, and the corresponding models were later evaluated on the test and holdout sets by taking the mean and standard deviation of their F1-scores. For MLPs, LSTMs, and GRUs, the validation fold was also used for early stopping during the training.

VI. RESULTS

A. Multi-class Classification

Addressing Q1, Table I shows the mean F1-score and standard deviation on the test set and, in parenthesis, on the holdout set using multi-class classification. With few exceptions on test and holdout sets, the sequential and non-sequential ML models tended to outperform the Feature-Based Heuristic. Table I does not include the Speech-and-Pause-Based Heuristic because this heuristic is applicable only to binary classification.

In general, the sequential and non-sequential models had similar F-1 performance in Table 1. For the test set, the results were mixed, with the RF and GRU having slightly better performance than other models for 2 participants, the MLP being slightly better for 3 participants, and the GRU being slightly better for all sessions. For the holdout set, the RF slightly outperformed other models.

TABLE I

MEAN F1-SCORE AND STANDARD DEVIATION ON THE TEST SET, AND, IN PARENTHESES, ON HOLDOUT SET, FOR THE **MULTI-CLASS CLASSIFICATION**, SEPARATELY FOR THE DIFFERENT SESSIONS.

		2 participants	3 participants	all sessions
non-sequential	DT	0.52±0.04 (0.50±0.06)	0.50±0.05 (0.37±0.07)	0.50±0.02 (0.46±0.07)
	RF	**0.59±0.01** (0.54±0.02)	0.49±0.01 (0.45±0.07)	0.52±0.01 (0.46±0.03)
	SVM	0.49±0.01 (0.42±0.01)	0.49±0.03 (0.40±0.05)	0.44±0.01 (0.43±0.02)
	MLP	0.54±0.01 (0.38±0.07)	**0.53±0.06** (0.38±0.08)	0.51±0.04 (0.47±0.02)
sequential	LSTM	0.58±0.02 (0.41±0.06)	0.51±0.04 (0.42±0.05)	0.50±0.04 (0.47±0.04)
	GRU	**0.59±0.01** (0.45±0.02)	0.47±0.04 (0.44±0.06)	**0.54±0.00** (0.42±0.02)
	Feature-Based **Heuristic**	0.48±0.11 (0.40±0.04)	0.40±0.08 (0.41±0.07)	0.50±0.06 (0.43±0.04)

TABLE II

MEAN F1-SCORE ON THE TEST SET OF ALL SESSIONS FOR **TWO-STEP BINARY CLASSIFICATION**, SPB: SPEECH-AND-PAUSE BASED HEURISTIC.

"not appropriate" vs. "appropriate/needed"

DT	RF	SVM	MLP	LSTM	GRU	SPB heuristic
0.70	0.74	0.68	0.70	0.74	**0.75**	0.54

"appropriate" vs. "needed"

DT	RF	SVM	MLP	LSTM	GRU
0.58	**0.61**	0.59	0.57	0.57	0.58

B. Two-step Binary Classification

To allow for comparison to multi-class classification in Q2, the mean F1-score of two-step binary classification for the test set of all sessions is provided in Table II. For the "not appropriate" vs. "appropriate/needed" classification, learning-based methods outperformed the Speech-and-Pause-Based Heuristic. The best models for binary classification on "not appropriate" vs. "appropriate/needed" were the GRU for sequential models and RF for non-sequential models, with the GRU having a slightly higher F-1 score (0.75 vs. 0.74).

For the "appropriate" vs. "needed" classification in Table II, there was an accuracy drop for the models using sequential and non-sequential data. In this classification, RF achieved the highest accuracy (0.61 average F-1 score), followed by SVMs (0.59). No heuristic approach was applicable to this classification setting, so they are omitted in Table II.

C. Multi-class vs Two-step Binary Classification

In the previous sections, we reported the aggregate performance for Multi-class and Two-step Binary classifiers. To answer Q2, we compare these results considering each class category. We use the SPB Heuristic as a baseline. The results for multi-class classification (using RF), two-step binary classification (using RF), and the heuristic are provided in Tables V, III and IV, respectively.

TABLE III

MEAN F1-SCORE OF **EACH CLASS CATEGORY** ON THE TEST SET OF ALL SESSIONS USING **TWO-STEP BINARY CLASSIFICATION** - REPORTED FOR RF. RESULTS FOR "NOT APPROPRIATE VS. APPROPRIATE/NEEDED" ON THE LEFT AND FOR "APPROPRIATE VS. NEEDED" ON THE RIGHT.

Class label	F1-score
not appropriate	0.73±0.01
appropriate/needed	0.74±0.02
F1-score (macro)	0.74±0.02

Class label	F1-score
appropriate	0.65±0.03
needed	0.58±0.05
F1-score (macro)	0.61±0.03

TABLE IV

MEAN F1-SCORE OF **EACH CLASS CATEGORY** ("NOT APPROPRIATE VS. APPROPRIATE/NEEDED") ON THE TEST SET OF ALL SESSIONS USING **SPEECH-AND-PAUSE-BASED HEURISTIC**.

Class label	F1-score
not appropriate	0.72
appropriate/needed	0.37
F1-score (macro)	0.54

TABLE V

MEAN F1-SCORE OF **EACH CLASS CATEGORY** ("NOT APPROPRIATE VS. APPROPRIATE VS. NEEDED") ON THE TEST SET OF ALL SESSIONS USING **MULTI-CLASS CLASSIFICATION** - REPORTED FOR RF.

Class label	F1-score
not appropriate	0.64±0.02
appropriate	0.56±0.03
needed	0.36±0.04

Compared to the SPB Heuristic, two-step binary classification by the RF model performs better for each class. Even though the heuristic had a relatively high score in the "not appropriate" class, it performed poorly on determining when a topic change is "appropriate/needed". For this category, the two-step binary approach using RF obtained ∼37% higher accuracy than the heuristic approach (Two-step Binary RF Model: %0.74±0.02, SPB heuristic:%0.37).

Lastly, when we compare one-step (multi-class classification) and two-step approaches (two-step binary classification), both models showed high performances in classifying the "not appropriate" class. Although both methods used RF as a classifier, there was a notable performance drop for the "appropriate" vs "needed" classes when the problem was formulated with a single-step approach.

D. Classification Performance Using Sets of Features

To investigate Q3, we provide the accuracy results of each model using varying sets of input features instead of using the whole feature set. The results were obtained using the Two-step Binary classification given the promise of this approach explored in Q2. Results are presented as the mean F1-score and standard deviation on the test set of all sessions and the holdout set (in parenthesis) in Table VI and VII.

In the test set (Table VI), the performance of the acoustic and Top-20 features was slightly higher than using all the features, with the GRU performing the best (0.76 ± 0.1 for acoustic and Top-20 features vs. 0.75 ± 0.01 for all features). For the holdout set, the best results were obtained with all features and the acoustic features. In particular, the RF model had a slightly higher performance with all the features (0.79 ± 0.01) than several other models with acoustic features (which reached 0.78 average F-1 scores). Thus, the main takeaway from these results is that using only Kinect body features to distinguish between "not appropriate" and "appropriate/needed" is worse than incorporating other features into this prediction problem. Surprisingly, acoustic features often led to good performance in this classification task.

All the models showed a decrease while classifying "appropriate" vs. "needed" (Table VII) compared to classifying "not appropriate" vs. "appropriate/needed" (Table VI), with no apparent correlations using different sets of features. Finally, holdout set accuracies (in both tables) showed similar trends with all session test set performances.

VII. DISCUSSION

In this study, we investigated the effectiveness of machine learning models in topic change prediction using non-verbal features. Given the complexity of the task, instead of following heuristic-based decisions, we suggest the need for learning-based methods for robots to be capable of topic moderation. Accordingly, we evaluate various ML models using sequential and non-sequential inputs, and we provide further analysis using one-step multi-class and two-step binary classification techniques. Our findings suggest the applicability of using ML approaches for topic change in robot-facilitated discussions. They also show that using acoustic data or the most informative features can provide comparable results with the whole future set. This can guide future HRI research to simplify features used without compromising the prediction performances.

While exploring Q1 and Q2, we compared a sensible heuristic against varying ML models (Table III, IV, and V). Even though the heuristic method had relatively high accuracy in the "not appropriate" class, it performed poorly on determining when a topic change is "appropriate/needed." This highlights the lack of flexibility and effectiveness of rule-based methods compared to the learning models and further highlights the need for learning methods for this task. Additionally, relatively similar performances of ML models on the test and hold-out data show these models' robustness and generalization capabilities (Table I).

Another finding demonstrates the robustness of ML models on unseen data obtained from the analysis for Q3 investigating varying sets of features (Table VI and VII). Regarding the type of features, acoustic and Top-20 features were identified as promising choices. Kinect-derived features showed low overall performance and generalization issues. Their performance could be attributed to the higher dimensionality of the features compared to acoustic features or person-specificity. An additional reason for the low impact of Kinect features could be the discussion topics during the brainstorming sessions, which, in contrast with previous work on gestures and topic structure [45], [42], did not mainly involve spatial information, that could otherwise encourage using hand gestures. In addition, cultural differences are known to affect the use of gestures [53], which might have influenced the results for the Kinect features.

TABLE VI

MEAN F1-SCORE AND STANDARD DEVIATION ON THE ALL SESSIONS TEST SET, AND IN PARENTHESES ON HOLDOUT SET, FOR **"NOT APPROPRIATE" VS. "APPROPRIATE/NEEDED"** CLASSES, USING **TWO-STEP BINARY CLASSIFICATION**.

	Acoustic	**Kinect**	**Top-20**	**All**
DT	0.72±0.02 (0.78±0.03)	0.54±0.03 (0.54±0.04)	0.72±0.01 (0.75±0.02)	0.70±0.03 (0.77±0.02)
RF	0.74±0.02 (0.76±0.02)	0.58±0.02 (0.56±0.06)	0.74±0.01 (0.76±0.02)	0.74±0.02 (0.79±0.01)
SVM	0.73±0.01 (0.78±0.01)	0.54±0.01 (0.54±0.02)	0.72±0.01 (0.75±0.00)	0.68±0.01 (0.72±0.01)
MLP	0.73±0.01 (0.78±0.01)	0.56±0.02 (0.56±0.03)	0.72±0.02 (0.76±0.01)	0.70±0.03 (0.74±0.03)
LSTM	0.74±0.01 (0.78±0.02)	0.58±0.02 (0.51±0.05)	0.75±0.02 (0.75±0.01)	0.74±0.03 (0.74±0.02)
GRU	**0.76±0.01** (0.77±0.02)	0.61±0.02 (0.49±0.04)	**0.76±0.01** (0.76±0.01)	0.75±0.01 (0.74±0.03)

TABLE VII

MEAN F1-SCORE AND STANDARD DEVIATION ON THE ALL SESSIONS TEST SET, AND IN PARENTHESES ON HOLDOUT SET, FOR **"APPROPRIATE" VS. "NEEDED"** CLASSES, USING **TWO-STEP BINARY CLASSIFICATION**.

	Acoustic	**Kinect**	**Top-20**	**All**
DT	0.59±0.03 (0.54±0.04)	0.54±0.03 (0.51±0.03)	0.57±0.04 (0.47±0.06)	0.58±0.06 (0.53±0.06)
RF	0.60±0.01 (0.57±0.06)	0.54±0.03 (0.46±0.05)	0.57±0.04 (0.53±0.09)	**0.61±0.03** (0.50±0.05)
SVM	0.59±0.02 (0.58±0.03)	0.56±0.02 (0.47±0.02)	0.51±0.03 (0.55±0.02)	0.59±0.02 (0.50±0.03)
MLP	0.58±0.03 (0.57±0.05)	0.56±0.04 (0.47±0.03)	0.54±0.05 (0.50±0.09)	0.57±0.02 (0.45±0.05)
LSTM	**0.61±0.02** (0.57±0.06)	0.53±0.02 (0.49±0.05)	0.55±0.03 (0.44±0.03)	0.57±0.03 (0.50±0.04)
GRU	0.59±0.03 (0.54±0.04)	0.50±0.01 (0.42±0.08)	0.51±0.06 (0.44±0.04)	0.58±0.02 (0.45±0.05)

Considering the type of classification techniques explored for Q2, two-step binary classification reported higher accuracy than multi-class classification. This could be due to the chance level increase when the process was simplified to binary classification instead of making a prediction among three classes. Additionally, in binary classification between "not appropriate vs. appropriate/needed" obtained higher accuracy than "appropriate vs. needed". This suggests that the decision is easier for the robot when it is not expected to change the discussion topic (higher accuracy for "not appropriate vs. appropriate/needed" prediction). However, when adding the possibility to change the topic, the decision is harder (lower accuracy for "appropriate vs. needed").

Regarding the type of ML models, there was no clear advantage in using sequential approaches over non-sequential approaches. These findings indicate that the aggregated features could provide enough information and, while combined with simpler models, achieve comparable results without the complexity of the sequential method. This finding is especially interesting for HRI contexts as simpler models also have lower data requirements, beneficial given the cost and complexity of collecting human-robot interaction data. Nonetheless, more data could benefit the models, especially for using sequential approaches, given the high dimensionality of the input; thus, the amount of data available could have affected the performance of our experiments.

One of the main challenges of topic change in robot-facilitated discussion is benchmarking. Given the complexity of the task, there are no learning-based baselines to build upon or publicly available datasets. This motivated us to gather our own dataset, yet it has a limitation of the finite quantity of participants and interaction sessions we had at our disposal, combined with its imbalanced nature. Therefore, the aim of future research could be the collection of a more extensive and diverse dataset. Lastly, future research could investigate how the proposed methods can be applied to larger groups, which could reveal additional opportunities and contribute to a wider social context and practical applications of robot-moderated discussions.

VIII. CONCLUSION

In this paper, we proposed a novel technical problem: when should a robot change the topic of a robot-moderated group discussion? Further, we provided insight into using machine learning approaches to solve this problem. Our results demonstrate the complexity of the task. Heuristic-based approaches under-performed learning-based methods, showing the value of machine learning in this problem domain. Our findings demonstrate the importance of selecting the most informative features when predicting topic changes and, surprisingly, suggested that acoustic features are particularly useful. Overall, our work provided a new dataset for automated topic change decisions in robot-moderated group discussions and an initial exploration of models that could be used to address this technical challenge.

REFERENCES

[1] D. R. Forsyth, *Group dynamics*. Wadsworth Cengage Learning, 2014.

[2] S. Sebo, B. Stoll, B. Scassellati, and M. F. Jung, "Robots in Groups and Teams: A Literature Review," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1–36, Oct. 2020.

[3] S. Gillet, M. Vázquez, S. Andrist, I. Leite, and S. Sebo, "Interaction-Shaping Robotics: Robots that Influence Interactions between Other Agents," *J. Hum.-Robot Interact.*, 2024.

[4] S. Sabanovic, C. C. Bennett, Wan-Ling Chang, and L. Huber, "PARO robot affects diverse interaction modalities in group sensory therapy for older adults with dementia," in *2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR)*. IEEE, 2013.

[5] C. Thompson, S. Mohamed, W.-Y. G. Louie, J. C. He, J. Li, and G. Nejat, "The robot Tangy facilitating Trivia games: A team-based user-study with long-term care residents," in *2017 IEEE International Symposium on Robotics and Intelligent Sensors (IRIS)*. IEEE, 2017.

[6] H. Tennent, S. Shen, and M. Jung, "Micbot: A Peripheral Robotic Object to Shape Conversational Dynamics and Team Performance," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Daegu, Korea (South): IEEE, 2019.

[7] S. Strohkorb, E. Fukuto, N. Warren, C. Taylor, B. Berry, and B. Scassellati, "Improving human-human collaboration between children with a social robot," *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016*, 2016.

[8] E. Short and M. J. Mataric, "Robot moderation of a collaborative game: Towards socially assistive robotics in group interactions," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2017.

[9] A. de Rooij, S. van den Broek, M. Bouw, and J. de Wit, "Co-designing with a social robot facilitator: Effects of robot mood expression on human group dynamics," in *Proceedings of the 11th International Conference on Human-Agent Interaction*, 2023.

[10] J. Geerts, J. de Wit, and A. de Rooij, "Brainstorming With a Social Robot Facilitator: Better Than Human Facilitation Due to Reduced Evaluation Apprehension?" *Frontiers in Robotics and AI*, vol. 8, 2021.

[11] E. Mizrahi, N. Danzig, and G. Gordon, "vrobotator: A virtual robot facilitator of small group discussions for k-12," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, 2022.

[12] G. Kovacs, T. Grosz, and T. Varadi, "Topical unit classification using deep neural nets and probabilistic sampling," in *7th IEEE International Conference on Cognitive Infocommunications*. IEEE, 2016.

[13] L. Hunyadi and I. Szekrényes, Eds., *The Temporal Structure of Multimodal Communication: Theory, Methods and Applications*, ser. Intelligent Systems Reference Library. Springer International Publishing, 2020, vol. 164.

[14] K. Tomiyama, F. Nihei, Y. I. Nakano, and Y. Takase, "Identifying Discourse Boundaries in Group Discussions using a Multimodal Embedding Space," in *ACM IUI 2018 Workshops, Symbiotic Interaction and Harmonious Collaboration for Wisdom Computing*, 2018.

[15] E. Schneiders, E. Cheon, J. Kjeldskov, M. Rehm, and M. B. Skov, "Non-dyadic interaction: A literature review of 15 years of human-robot interaction conference publications," *J. Hum.-Robot Interact.*, vol. 11, no. 2, 2022.

[16] M. F. Jung, N. Martelaro, and P. J. Hinds, "Using Robots to Moderate Team Conflict: The Case of Repairing Violations," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, 2015.

[17] S. Shen, P. Slovak, and M. F. Jung, ""Stop. I See a Conflict Happening."," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2018.

[18] H. Erel, D. Trayman, C. Levy, A. Manor, M. Mikulincer, and O. Zuckerman, "Enhancing emotional support: The effect of a robotic object on human–human support quality," *International Journal of Social Robotics*, pp. 1–20, 2021.

[19] S. Strohkorb Sebo, M. Traeger, M. F. Jung, and B. Scassellati, "The Ripple Effects of Vulnerability: The Effects of a Robot's Vulnerable Behavior on Trust in Human-Robot Teams," *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction - HRI '18*, no. February, pp. 178–186, 2018.

[20] S. Strohkorb Sebo, L. L. Dong, N. Chang, and B. Scassellati, "Strategies for the Inclusion of Human Members within Human-Robot Teams," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 3 2020.

[21] S. Gillet, W. van den Bos, and I. Leite, "A social robot mediator to foster collaboration and inclusion among children," in *Proceedings of Robotics: Science and Systems*, Corvalis, Oregon, USA, 2020.

[22] S. Tuncer, S. Gillet, and I. Leite, "Robot-mediated inclusive processes in groups of children: From gaze aversion to mutual smiling gaze," *Frontiers in Robotics and AI*, vol. 9, 2022.

[23] S. Gillet, R. Cumbal, A. Pereira, J. Lopes, O. Engwall, and I. Leite, "Robot Gaze Can Mediate Participation Imbalance in Groups with Different Skill Levels," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2021.

[24] M. Vázquez, E. J. Carter, J. A. Vaz, J. Forlizzi, A. Steinfeld, and S. E. Hudson, "Social group interactions in a role-playing game," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, 2015.

[25] M. Vázquez, E. J. Carter, B. McDorman, J. Forlizzi, A. Steinfeld, and S. E. Hudson, "Towards robot autonomy in group conversations: Understanding the effects of body orientation and gaze," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017.

[26] L. Grassi, C. T. Recchiuto, and A. Sgorbissa, "Knowledge-Grounded Dialogue Flow Management for Social Robots and Conversational Agents," *International Journal of Social Robotics*, vol. 14, no. 5, 2022.

[27] N. Glas and C. Pelachaud, "Topic management for an engaging conversational agent," *International Journal of Human-Computer Studies*, vol. 120, 2018.

[28] J. Tang, T. Zhao, C. Xiong, X. Liang, E. Xing, and Z. Hu, "Target-Guided Open-Domain Conversation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019.

[29] K. Sevegnani, D. M. Howcroft, I. Konstas, and V. Rieser, "OTTers: One-turn Topic Transitions for Open-Domain Dialogue," in *Proceedings of the 59th Ann. Meeting of the Assoc. for Computational Linguistics and the 11th Int. Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 2021.

[30] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human-robot interaction," in *5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2010.

[31] C. Peters, "Direction of Attention Perception for Conversation Initiation in Virtual Environments," in *Intelligent Virtual Agents*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, vol. 3661, series Title: Lecture Notes in Computer Science.

[32] R. Ishii, Y. I. Nakano, and T. Nishida, "Gaze awareness in conversational agents: Estimating a user's conversational engagement from eye gaze," *ACM Transactions on Interactive Intelligent Systems*, vol. 3, no. 2, 2013.

[33] J. Eisenstein, R. Barzilay, and R. Davis, "Gestural Cohesion for Topic Segmentation," in *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics, 2008.

[34] S. Nakajima and J. F. Allen, "A Study on Prosody and Discourse Structure in Cooperative Dialogues," *Phonetica*, vol. 50, no. 3, 1993.

[35] R. Herman, "Phonetic markers of global discourse structures in English," *Journal of Phonetics*, vol. 28, no. 4, 2000.

[36] M. Swerts, D. G. Bouwhuis, and R. Collier, "Melodic cues to the perceived "finality" of utterances," *The Journal of the Acoustical Society of America*, vol. 96, no. 4, 1994.

[37] M. K. Zellers, "Prosodic Detail and Topic Structure in Discourse," Ph.D. dissertation, University of Cambridge UK, 2011.

[38] F. Quek, Y. Xiong, and D. McNeill, "Gestural trajectory symmetries and discourse segmentation," in *7th International Conference on Spoken Language Processing (ICSLP 2002)*. ISCA, 2002.

[39] S. Duncan, "Some signals and rules for taking speaking turns in conversations." *Journal of Personality and Social Psychology*, vol. 23, no. 2, 1972.

[40] M. Zellers, D. House, and S. Alexanderson, "Prosody and hand gesture at turn boundaries in Swedish," in *Speech Prosody 2016*. ISCA, 2016.

[41] J. Streeck and U. Hartge, "Previews: Gestures at the Transition Place," in *Pragmatics & Beyond New Series*, P. Auer and A. Di Luzio, Eds. John Benjamins Publishing Company, 1992, vol. 22.

[42] J. Cassell, Y. I. Nakano, T. W. Bickmore, C. L. Sidner, and C. Rich, "Non-verbal cues for discourse structure," in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01*. Association for Computational Linguistics, 2001.

[43] C. Goodwin and M. H. Goodwin, "Concurrent Operations on Talk: Notes on the Interactive Organization of Assesments," *IPrA papers in pragmatics*, vol. 1, no. 1, pp. 1–54, Jan. 1987.

[44] F. Rossano, "Gaze in Conversation," in *The Handbook of Conversation Analysis*, 1st ed., J. Sidnell and T. Stivers, Eds. Wiley, 2012.

[45] F. Quek, D. McNeill, R. Bryll, C. Kirbas, H. Arslan, K. McCullough, N. Furuyama, and R. Ansari, "Gesture, speech, and gaze cues for discourse segmentation," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, vol. 2. IEEE Comput. Soc, 2000.

[46] S. Thompson, A. Narcomey, A. Lew, and M. Vázquez, "Shutter: A Low-Cost and Flexible Social Robot Platform for In-the-Wild Deployments," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '24. Association for Computing Machinery, 2024.

[47] G. Skantze, "Turn-taking in Conversational Systems and Human-Robot Interaction: A Review," *Computer Speech & Language*, vol. 67, 2021.

[48] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, vol. 25, no. 3, 2011.

[49] M. Johansson and G. Skantze, "Opportunities and Obligations to Take Turns in Collaborative Multi-Party Human-Robot Interaction," in *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Prague, Czech Republic: Association for Computational Linguistics, 2015.

[50] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML'10. Omnipress, 2010.

[51] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *3rd International Conference on Learning Representations, ICLR 2015 Conference Track Proceedings*, 2015.

[52] F. Ferri, P. Pudil, M. Hatef, and J. Kittler, "Comparative study of techniques for large-scale feature selection," in *Machine Intelligence and Pattern Recognition*. Elsevier, 1994, vol. 16.

[53] A. Kendon, "Geography of gesture," *Semiotica*, vol. 37, no. 1/2, pp. 129–163, 1981.