Using Valence Emotion to Predict Group Cohesion's Dynamics: Top-down and Bottom-up Approaches

Lucien Maman
LTCI, Télécom Paris
Institut Polytechnique de Paris
Palaiseau, France
lucien.maman@telecom-paris.fr

Mohamed Chetouani
CNRS-ISIR
Sorbonne University
Paris, France
mohamed.chetouani@sorbonne-universite.fr

Laurence Likforman-Sulem
LTCI, Télécom Paris
Institut Polytechnique de Paris
Palaiseau, France
laurence.likforman@telecom-paris.fr

Giovanna Varni
LTCI, Télécom Paris
Institut Polytechnique de Paris
Palaiseau, France
giovanna.varni@telecom-paris.fr

***Abstract*—Cohesion is an affective group phenomenon. It has received a lot of attention from scholars both in Social Sciences and in Affective Computing that showed that cohesion and emotion influence each other, highlighting the need to jointly analyze them. This study presents 2 deep neural network architectures grounded on multitask learning to jointly predict cohesion and emotion. Inspired by 2 major Social Sciences approaches on group emotion (i.e., Top-down and Bottom-up), these architectures exploit cohesion and emotion interdependencies intending to improve the prediction of the dynamics (i.e. changes over time) of the Social and Task dimensions of cohesion. Emotion, here, is addressed in terms of its valence. Both architectures are evaluated against the performances of a similar model that only predicts the dynamics of both the Social and Task dimensions of cohesion, without integrating valence. Statistical analysis shows that only the deep model implementing the Bottom-up approach significantly improved the predictions of the Task cohesion's dynamics. This result confirms the theoretical and practical benefits of multitasking as it takes full advantage of the inherent relationships between group emotion and cohesion to improve Task cohesion's predictions.**

***Index Terms*—Group Cohesion, Group Dynamics, Group Emotion, Multimodal Interaction, Multitask Learning**

I. Introduction

Everyday life mainly happens in groups. Social interaction elicits emotions that can, for example, shape groups [1] and influence judgments and choices [2]. Although its relevance in Affective Computing, to date, affect and emotion in groups have been under-investigated even though they are potential drivers of emergent states. They are group phenomena resulting from the micro-level affective, behavioral and cognitive interactions among group members, through the micro-processes of group interaction (e.g., [3]). According to [4], affective emergent states reflect the relationships among group members and their emotional responses. Cohesion is the affective emergent state that has received the most attention from scholars both in Social Sciences and in Affective Computing [5]–[11]. It is a multidimensional construct that is the result of all the forces acting on members to remain in the group [12]. Several definitions of cohesion, however, exist (e.g., [13]–[15]) and they attribute different dimensions to this emergent state. In all the definitions the Social and Task dimensions are always mentioned and relate to the aspects that highlight the goal- and task-based activities of the group. As Severt and Estrada state in [16], relationships between cohesion and other constructs exist (e.g., group trust and leadership) but may vary depending on the relationships between the members of the group (e.g., hierarchical relationships) and the dimension in which cohesion is investigated. Moreover, the link between cohesion and individual and group emotion has been particularly studied in Social Sciences (e.g., [17]), showing that cohesion and emotions influence each other [18], [19]. For example, highly cohesive teams likely promote positive emotions such as happiness among group members. Reciprocally, positive individuals likely create a climate conducive to cohesion. Building upon evidence showing that positively or negatively valenced emotions could affect cohesion in tasks requiring group decision making or creativity [20], emotion is here addressed in terms of its valence.

In this paper, we present 2 deep neural network architectures that use multitask learning to jointly predict cohesion's dynamics (i.e., increase or decrease) and emotion, exploiting the interdependencies between these phenomena. They are inspired by the Top-down and Bottom-up approaches from the Social Sciences [17]. Top-down considers the group's dynamic processes (e.g., emergent states such as cohesion) as responsible for a group level emotion that influences the members' feelings and behavior. Oppositely, the Bottom-up approach views group-level emotion as the sum of its individuals' affective compositions [17]. To assess whether

This paper has been partially supported by the French National Research Agency (ANR) in the framework of its Technological Research JCJC program (GRACE, project ANR-18-CE33-0003-01, funded under the Artificial Intelligence Plan).

and how predicting valence improves the prediction of the cohesion's dynamics for these dimensions, both architectures are evaluated against a model predicting the dynamics of the Social and Task dimensions of cohesion in a multilabel setting, without integrating the cohesion-emotion relationships.

II. Background and Related work

A. Background

1) Cohesion: In the 1940s, Lewin first defined cohesion as *"a group characteristic that depends on its size, organization and intimacy"* [21]. Following Lewin's work, Festinger referred to cohesion as the *"total field of forces causing members to remain in the group"* [12]. Then, the notion of cohesion evolved from a single to a multidimensional construct. Recently, Severt and Estrada proposed an integrative framework of cohesion [16]. According to it, cohesion is structured into functional properties. In this paper, we focus on the instrumental property, which comprises the Social and Task dimensions as they both play a role in a wide range of situations and are relevant for studying task-driven groups. It is important to note that the Social dimension refers to the social bonds between group members that are bound by the group's working relationship while the Task dimension relates to the degree of commitment to group tasks and goals [16].

2) Emotions in groups: Emotions can either bind or splinter a group [22]. Barsade and Gibson highlighted 2 approaches to characterize group emotions [17]. *Top-down* focuses on the group as a whole. This means that group dynamics influence the feelings and behaviors of members of the group. Following this approach, scholars characterized group emotions as (1) forces which shape individual emotional response (e.g., [23]), (2) social norms (e.g., [24]), (3) the interpersonal glue that keeps groups together (e.g., [12]) and (4) a display of group's maturity and development (e.g., [25]). In this study, we follow the first characterization of group emotion. *Bottom-up* investigates how the emotions of group members combine to create a group-level emotion, approximating the group as the sum of its parts. This approach led researchers to examine the group through a variety of compositional perspectives such as the mean of the group's members, the degree of emotional variance within the group and the influence of the most emotionally extreme members of the group. There is, however, an open debate on defining the best approach. As both the Top-down and the Bottom-up approaches bring different characterizations of group emotion, Barsade and Gibson recommend exploring methods following both these approaches to have a complete picture of this phenomenon [17]. Literature on cohesion and group emotion highlighted the importance to consider them at both individual and group levels [26].

B. Related work

As the aim of the paper is to explore whether and how integrating the cohesion-emotion interplay could improve cohesion's dynamics prediction, this Section focuses on automated studies addressing cohesion only or both cohesion and emotion. First studies in Computer Science addressing cohesion were aimed at predicting cohesion levels (i.e., low and high) in diverse contexts (e.g., meetings) within small groups interactions. The works focusing on nonverbal cues usually yield better results than the ones focusing on verbal communication only [27]. Furthermore, studies using a multimodal approach obtained better performances compared to the ones using unimodal approaches (e.g., [28], [29]). For example, Hung and Gatica-Perez investigated Social and Task dimensions of cohesion in meetings, using external assessment of cohesion and multimodal audio-visual features [6]. Nanninga and colleagues extended this work, integrating pairwise and group features related to the alignment of para-linguistic speech behavior [7]. Both studies addressed cohesion prediction as a binary classification problem but the latter proposed a pair of distinct binary classifiers for the Social and Task dimensions, respectively instead of a single model predicting high and low levels of cohesion. Recently, Walocha *et al.*, developed a model that jointly predicts the dynamics of the Social and Task dimensions of cohesion in a multilabel setting to inspect the relationships between these dimensions [8]. All these studies, however, neglected the relationships of cohesion with other phenomena such as emotion. More specifically, they did not investigate how the prediction of cohesion could benefit from this one. Latterly, Dhall et al., provided a benchmarking platform to investigate newer problems related to Affective Computing within the context of the EmotiW challenge [9]. In this challenge, researchers implemented various methods to predict group cohesion from images (e.g., [30], [31]) but also to jointly predict emotion and cohesion's level in images and videos using deep learning networks. Zou and colleagues, for example, presented a hybrid deep learning network for the prediction of group emotion and level of cohesion [10]. They first used a model to classify emotions according to their valence (positive, neutral, negative) and used the model's output into a regression layer to predict the cohesion level (between 0 and 3). They also implemented a multitask loss to merge the regression task (i.e., the prediction of the level of cohesion) with the classification task (emotion prediction). They reached an accuracy rate of classification of 74.8% for the prediction of the valence of emotion, and a Mean Squared Error (MSE) of 0.7 for their cohesion regression task. Also, in [11], the authors designed a multimodal deep neural network based on the inception V3 pre-trained model [32] to jointly predict group emotion and cohesion in videos. The videos used in this study largely differed, for example, in scenarios and poses, making it difficult for the model to capture the dynamics of group emotions since it was trained on images. Their model predicts the valence of emotions (i.e., positive, neutral, negative) with 47.5% accuracy and predicts cohesion with 0.8 MSE across the test set. In contrast to these studies, we investigate how emotion can be related to a specific dimension of cohesion. We also explore various approaches of group emotion (e.g. Top down, Bottom up) to improve the joint prediction of cohesion and emotion.

III. Computational framework and Experimental settings

Building upon this knowledge, we conceived a computational framework for jointly studying cohesion and emotions in small group interactions (see Fig. 1). The framework takes into account multimodal human behaviors both at individual and group levels through handcrafted or deep features. These are the input to predict the dynamics of cohesion. In addition, the relationships between cohesion and emotion are either modeled following the Bottom-up or the Top-down approach.

A. Dataset

GAME-ON is a multimodal dataset consisting of more than 11 hours of small group interactions in which no roles were attributed to the participants [33]. The interactions are organized in 5 different tasks eliciting variations of cohesion (see Table I). The dataset is composed of data from 15 groups of 3 persons. For each group, audio, video and motion capture recordings are available as well as self-assessments of cohesion and emotion. The average duration of each group session is 35min 30s (SD = 4min 10s). In our study, we use the motion capture data (captured at 100 Hz) and the audio data (captured at 48 kHz). Social and Task dimensions of cohesion were assessed through a slightly modified version of the Group Environment Questionnaire (GEQ) [34]. It includes items concerning both individuals (7 items) and the group as a whole (7 items), making it suitable for both Bottom-Up and Top-down approaches. About emotion assessment, participants could choose, between each task, which among 6 emotions (i.e., *Angry*, *Frustrated*, *Ashamed*, *Proud*, *Admiring*, *Happy* following [35]) better reflected their feelings.

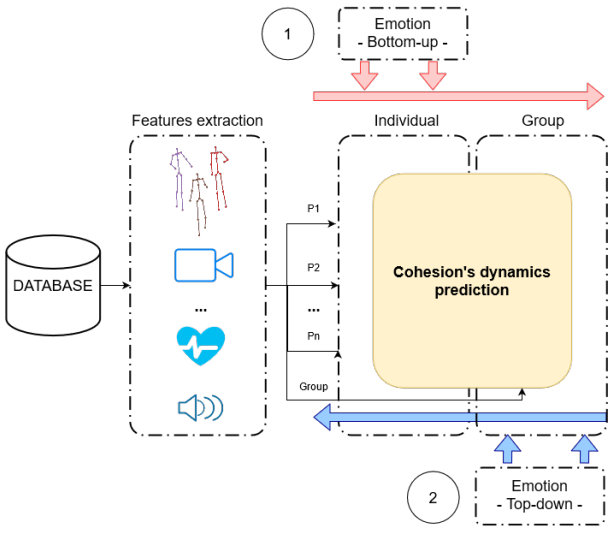


Fig. 1. The framework designed for jointly studying cohesion and emotion. It handles the integration of group emotion following the Bottom-up or the Top-down approaches. Multimodal features are extracted and feed the cohesion module to predict the dynamics of cohesion.

TABLE I
Variations of cohesion between two consecutive tasks (T) for the Social and Task dimensions. Start means the beginning of the recordings.

	Variation of cohesion	
Transition	**Social**	**Task**
Start - T1	Decrease	Decrease
T1 - T2	Decrease	Increase
T2 - T3	Increase	Decrease
T3 - T4	Increase	Increase
T4 - T5	Increase	Increase

B. Handcrafted features

We developed and extracted 84 motion capture-based and audio nonverbal features characterizing social interaction. For the sake of brevity and narrative clarity, implementation details are not given here. A thin slices approach was adopted and analysis was carried out over consecutive time windows. This approach refers to the process of making very quick inferences about the individual and/or group phenomena with a minimal amount of information [36]. A duration of 20s for the time windows was chosen according to previous work on group interaction [37] and cohesion perception [38]. Some of the features are computed by applying statistical functions on the behavioral descriptors listed in Table II (see the ones with a '$\star$'), according to their modality. For others, the raw values of the descriptors are retained. The Table also reports how each feature is computed and whether each descriptor is computed at the individual or group level. It is worth mentioning that, for clarity reasons, some of the names chosen for describing the features reported in Table II concern an ensemble of features that are related to the same behavior (e.g., posture expansion regroups both the latitudinal and longitudinal expansion features). All of the features extracted were either used in previous computational studies investigating cohesion (e.g., [6]–[8]) or identified as relevant from Social Sciences studies on cohesion (e.g., synchrony [39]). In short, motion capture features are related to proxemics and kinesics as the way people use the space (proxemics) as well as their body movement and gesture (kinesics) play an important role in nonverbal communication [40]. Concerning the audio features, the Geneva Minimalistic Acoustic Parameter Set (GeMAPS; see [41]) was selected. Moreover, turn-taking-based features were also included.

C. Labels

The dynamics of cohesion and the valence of group emotion are addressed as binary classification problems (i.e., Increase/Decrease and Positive/Negative, respectively).

1) Cohesion: To build labels for cohesion, for each pair of consecutive tasks (e.g., the second task and the third one) and for each dimension (i.e., Social and Task), we ranked the scores given by the 3 group members and we took the mean of the rank differences of each player. This strategy allowed us to approximate the changes of cohesion that occurred during the interaction while limiting the impact of the variability of the

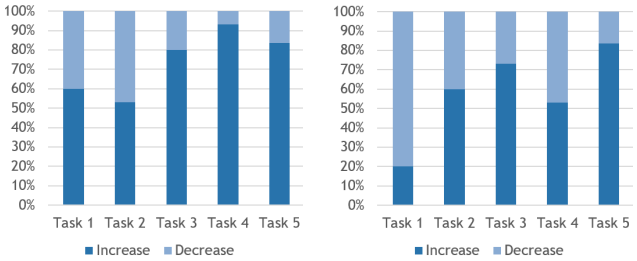
TABLE II

LIST OF THE MOTION CAPTURE-BASED AND AUDIO NONVERBAL FEATURES CHARACTERIZING SOCIAL INTERACTION USED IN THIS STUDY. THE FEATURES WITH A '$\star$' ARE THE ONES FOR WHICH WE APPLIED STATISTICAL FUNCTIONS (I.E., MEAN, STD, MIN, MAX AND SKEWNESS).

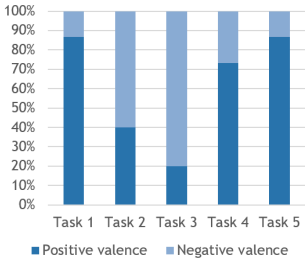
	Motion Capture		**Audio**	
	Proxemics	**Kinesics**	**GeMAPS**	**Turn-taking**
Individual	Distance from individual to the barycentre of the group$\star$ Total distance traveled	Posture expansion$\star$ Kinetic energy$\star$	Pitch / Jitter / Loudness Spectral slope / Harmonic differences F1, F2, F3 frequency and relative energy F1 bandwidth	Laughter duration Total speaking time
Group	Maximum distance between group members$\star$ Time in F-Formation (triangle or semi circular) Interpersonnal distances (Public, Social and Personnal spaces)	Amount of walking$\star$ Amount of hand gesture while not walking$\star$ Touch detection$\star$ Synchrony of kinetic energies		Average turn duration Time of overlapping speech

group members' ratings introduced by the self-assessments. Finally, we binarized these means of the rank differences as $\{0, 1\}$ based on their sign. In particular, a value equal to 0 is assigned when the mean is negative (i.e. a decrease in cohesion occurred), a value equal to 1 is assigned when the mean is positive (i.e. no change or an increase in cohesion occurred). In this work, we explicitly focused on decreases in cohesion. This, indeed, is an established method in research on Affective Computing (e.g., [42]). The labels' distribution is imbalanced for the Social dimension (73% of "Increase" labels), whereas it is balanced for the Task dimension (56% of "Increase" labels). Figure 2(b) shows the labels' distribution per task. Labels are highly imbalanced for most of the tasks. This is, however, expected since the GAME-ON dataset was conceived to explicitly control the dynamics of cohesion.

2) Valence: As previously mentioned, we focused on the valence of emotion as it has a direct impact on cohesion, independently of the approach chosen (i.e., Top-down or Bottom-up) [17], [18]. Valence labels are obtained in the following way. We first assigned a valence (positive or negative) to every emotion picked up by each group member after each task (more than one emotion could be provided per group member), according to [35]. Then, for each task and each group, we summed up $+1$ if a group member chose an emotion with a positive valence (e.g., happy) or -1 if a group member chose an emotion with a negative valence (e.g., ashamed). Depending on the sign of this sum, we defined the group valence as positive or negative. This labelling strategy resulted in a slightly imbalanced distribution (61% of positive emotion). Similarly to the cohesion labels, high imbalances for each task occurred (see Fig. 2(c)).



(a) Social dimension of cohesion (b) Task dimension of cohesion



(c) Valence of emotion

Fig. 2. Labels distribution of the 5 tasks for the Social and Task dimensions of cohesion (see 2(a) and 2(b)) and for the valence of emotion (see 2(c)).

IV. DEEP ARCHITECTURES

A. Modeling cohesion

To model the dynamics of the Social and Task dimensions of cohesion, we designed the "*from Individual to Group*" (fItG) architecture (see Fig. 3). It is inspired by the Team LSTM model developed by Kasparova *et al* [43] that learns specific patterns to predict student engagement, using individual unimodal features. FItG uses both multimodal individual and group features to learn a higher common representation, merging individual and group representations to predict cohesion. It consists of 4 modules. The *Feature Extraction module* extracts the handcrafted features described in Section III-B. The architecture takes as input the 2 last minutes of each task, resulting in a total of 30 thin slices of 20s per group. This choice is due to the fact that the cohesion scores used to produce the labels were likely influenced by the cohesion perceived around the end of each task. This effect was observed in many studies adopting different questionnaires (e.g., leader behavior assessment [44]). Due to the relatively small size of the dataset, the module performs data augmentation by creating synthetic groups. It is done by permuting the order of the group members of each group, resulting in a dataset 6 times bigger. The features of each group member are the input of the *Individual module* while the group features concur to make the input of the *Group module*. The Individual module is made of 3 branches, each one composed of a fully connected layer (FC layer) with a ReLu activation function and 50 units, followed by an LSTM layer. This structure enables the integration of temporality between the segments of the interaction. This module aims at learning a higher-level representation of an individual from the individual features. The model might learn

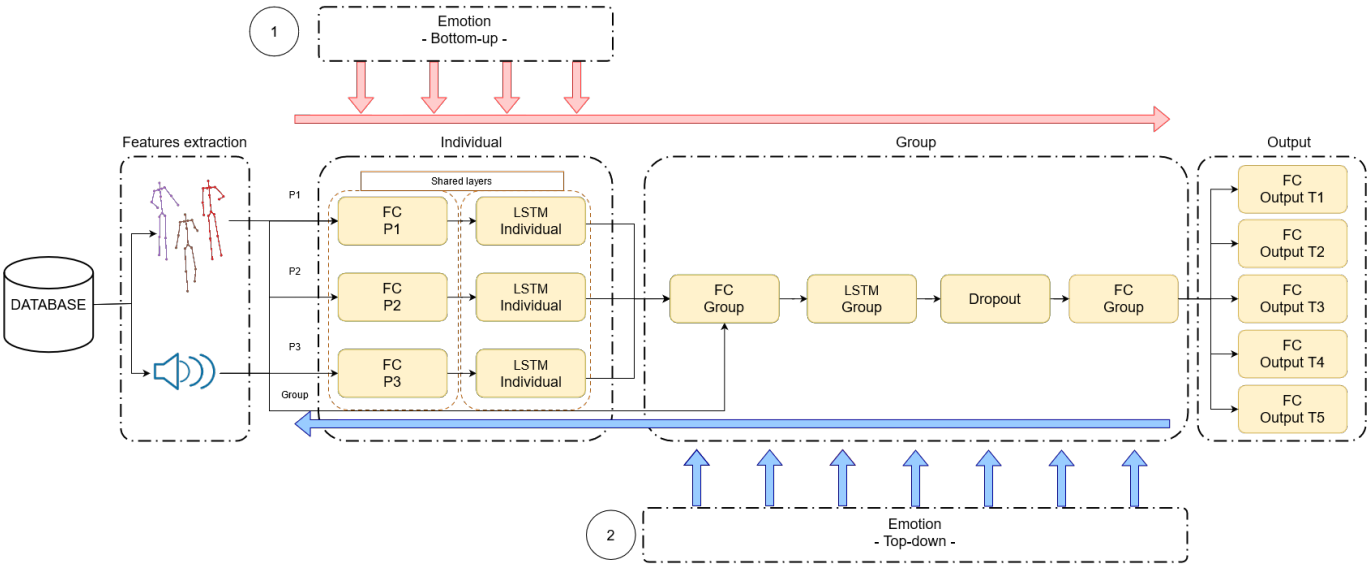


Fig. 3. The architecture of the fItG model. It is composed of 4 modules (i.e., *Features extraction*, *Individual*, *Group* and *Output*). Interaction is modeled both at individual and group levels. The dynamics of cohesion are predicted for the 5 tasks in a multilabel setting. The architecture is composed of fully connected (FC), LSTMs and Dropout layers.

undesired patterns related to the order in which each individual is processed by it across the multiple groups in the training set (e.g., learning a pattern specific to all the first group members seen by the model). To avoid this issue, we shared each layer of the 3 individual branches of the *Individual* module (i.e., the FC and LSTM layers). A common representation is learnt, for each layer, as:

$$\boldsymbol{Y}_i = \phi(\sum_{j=1}^{n}(W\boldsymbol{X}_j)) \tag{1}$$

where Y_i is the output of layer i, ϕ_i, the activation function of the layer i, W, the matrix of parameters common to every group members and X_j, the input related to player j. As groups are composed of 3 persons, n was here set equal to 3. The outputs of the 3 individual LSTM layers from the Individual module are then concatenated with the group features as input of the Group module. This module is aimed at learning temporal dynamics of cohesion at the group level. The module is made of a first FC layer with a ReLu activation function and 64 units, followed by an LSTM layer to integrate the group temporality. Next, a Dropout layer with a rate of 0.2 is used to prevent the model from overfitting. This is followed by another FC layer with a ReLu activation function and 16 units. Finally, the *Output module* consists of a FC layer with a sigmoid activation function and 2 units, for each task, predicting the Social and Task dimensions of cohesion in a multilabel setting.

B. Modeling valence

To implement the Bottom-up and the Top-down approaches, emotion was integrated into the fItG model using multitask learning. We took inspiration from [45] that designed a framework to jointly predict arousal, valence and dominance using multitask learning. They proved that a primary task (i.e., predicting arousal) could benefit from multitask learning by taking advantage of the shared representation of the features jointly learnt with the secondary tasks (i.e., predicting valence and dominance). Similarly, according to [18] stating that relationships exist between emotion and the Social and the Task dimensions, we expect that the prediction of the dynamics of these cohesion's dimensions (taken as the primary task) will be improved by the knowledge extracted from the prediction of valence (taken as the secondary task).

1) Bottom-up - the fIgT_Bu architecture: To integrate valence following the Bottom-up approach, we designed the fIgT_Bu architecture. The 3 combined outputs of the individual LSTMs from the Individual module are taken as input. This input feeds 2 FC layers with a ReLu activation function and 64 and 16 units, respectively. These layers are followed by a FC layer with a sigmoid activation function for each task and 1 unit. These final layers predict the valence of group emotion for each task (see Fig. 4). As valence is predicted from the output of the Individual module of the fItG, it has, during training, a direct impact on the common shared representation of an individual. The Individual module being part of the input of the Group module, integrating emotion following the Bottom-up approach also affects the group representation.

2) Top-down - the fItG_Td architecture: The fItG_Td architecture was designed to implement the Top-down approach. This model is characterized as depicted in Fig. 5. The output of the Group module is taken as input. A FC layer with a sigmoid activation function and 1 unit for each of the 5 tasks is used. In this way, the group and individual representations will both be impacted by the valence prediction during backpropagation. Finally, as we are considering these problems as binary classifications, a binary cross-entropy loss combined with the Adam optimizer was applied to each output.

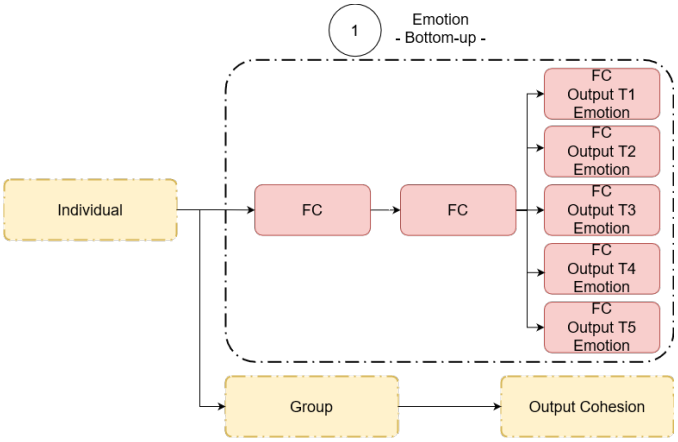


Fig. 4. The fItG_Bu architecture. The output of the individual module of the fItG model (i.e., yellow boxes) is processed into 2 FC layers with a ReLu activation function, followed by a FC layer with 1 unit and a sigmoid activation function, outputting the valence of emotion for each tasks.

V. Experimental evaluation

We evaluated the fItG_Bu and the fItG_Td architectures vs. the fItG architecture through a Leave-One-Group-Out (LOGO) cross-validation. We followed Colas *et al.*'s guidelines [46] to obtain a more robust measure of the architectures performances. We trained our models on 15 different randomly extracted seeds and averaged the performances. In this way, we aim at providing a reliable assessment of the models' performances. For each seed, we adopted a grid-search approach to select the learning rate (in $\{0.01, 0.001, 0.0001\}$) and the number of epochs (in $\{50,100, 200, 300, 500, 1000\}$). The weighted F1-score was chosen as it accounts for the label imbalance. Possible differences in the architectures' performances were assessed via computationally-intensive randomization tests. These are non-parametric tests avoiding the independence assumption between the results being compared and are suitable for non-linear measures such as F1-score [47].

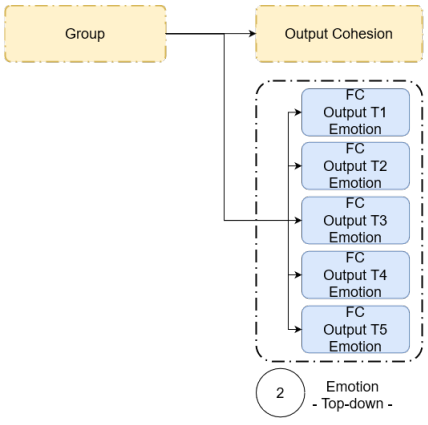
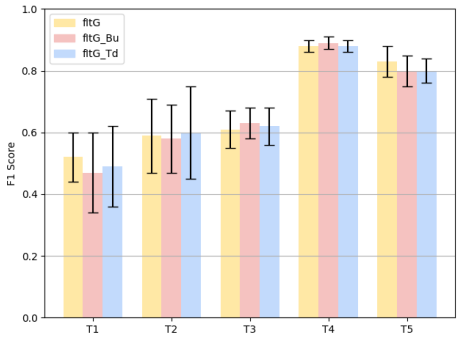


Fig. 5. The fItG_Td architecture. The output of the group module of the fItG model (i.e., yellow boxes) is processed into a FC layer with a sigmoid activation function and 1 unit, outputting the valence of emotion for each task.

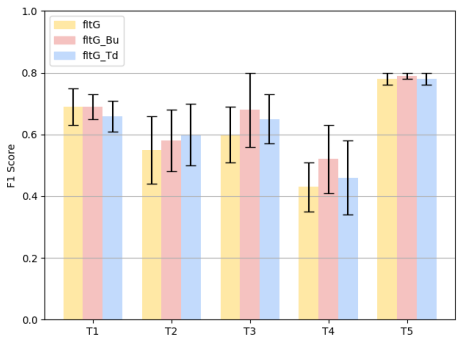
VI. Results and discussion

Table III summarizes the performances of the 3 architectures, in terms of F1-score. The fItG model obtained an averaged (over the tasks) weighted F1-score of 0.69 $\pm$0.03 for the Social dimension and 0.61 $\pm$0.03 for the Task dimension. As displayed by Fig. 6, T1 for the Social dimension and T4 for the Task dimension are particularly miss-predicted. The difficulties of the model to predict T1 could be because the Social dimension of cohesion builds over time to create flexible and constructive relationships [16]. Concerning the Task dimension, these difficulties might come from the nature of the task where group members have to agree to a common solution to solve a quiz. When disagreeing, group members indeed might provide very different cohesion scores, making it harder for the model to predict the dynamics of cohesion. FItG_Bu reached an averaged weighted F1-score of 0.67 $\pm$0.03 for the Social dimension and 0.65 $\pm$0.04 for the Task dimension, while it achieved 0.65 $\pm$0.03 for valence. FItG_Td, meanwhile, obtained an averaged weighted F1-score of 0.68 $\pm$0.03, 0.63 $\pm$0.03 and 0.64 $\pm$0.04 for the Social and Task dimensions and for the valence, respectively.

To compare the performances obtained by the architectures, we chose a significance level of $\alpha = 0.05$ for the statistical tests. For the sake of brevity, only the significant results are detailed. The statistical tests show that a significant difference between the 3 architectures occurred for the Task



(a) Social cohesion



(b) Task cohesion

Fig. 6. Average weighted F1-score per task for Social (a) and Task (b) dimensions. FItG is in yellow, fItG_Bu in red, and fItG_Td in blue.

TABLE III
SUMMARY OF THE WEIGHTED F1-SCORES OF THE PRIMARY AND SECONDARY TASKS (PREDICTING COHESION'S DYNAMICS AND VALENCE OF EMOTION, RESPECTIVELY) PER TASK AND PER DIMENSION FOR THE FITG, THE FITG_BU AND THE FITG_TD MODELS.

	Weighted F1-scores ±SD							
	fItG		**fItG_Bu**			**fItG_Td**		
	Social	**Task**	**Social**	**Task**	**Valence of emotion**	**Social**	**Task**	**Valence of emotion**
T1	0.52 ±0.08	0.69 ±0.06	0.47 ±0.13	0.69 ±0.04	0.76 ±0.05	0.49 ±0.13	0.66 ±0.06	0.78 ±0.07
T2	0.59 ±0.12	0.55 ±0.11	0.58 ±0.11	0.58 ±0.10	0.55 ±0.10	0.60 ±0.15	0.60 ±0.10	0.40 ±0.13
T3	0.61 ±0.06	0.60 ±0.09	0.63 ±0.05	0.67 ±0.12	0.66 ±0.03	0.62 ±0.06	0.65 ±0.08	0.67 ±0.05
T4	0.88 ±0.03	0.43 ±0.08	0.88 ±0.02	0.52 ±0.10	0.47 ±0.08	0.88 ±0.02	0.46 ±0.12	0.57 ±0.08
T5	0.84 ±0.05	0.78 ±0.02	0.81 ±0.05	0.79 ±0.02	0.79 ±0.02	0.80 ±0.04	0.78 ±0.02	0.78 ±0.02
Average	**0.69 ±0.03**	**0.61 ±0.03**	**0.67 ±0.03**	**0.65 ±0.04**	**0.65 ±0.03**	**0.68 ±0.03**	**0.63 ±0.03**	**0.64 ±0.04**

dimension only ($p = .016$). A possible explanation is that positive emotions maintain a particularly strong relationship with social cohesion [18], making it more difficult for the model to differentiate these phenomena. A post-hoc analysis using pairwise permutation t-tests was carried out. These tests reveal that only fItG_Bu reaches significance ($p = .012$). This improvement in performance (from 0.61 ±0.03 to 0.65 ±0.04) indicates that integrating valence in a Bottom-up fashion helps the model to learn a better representation of an individual, leading to a more accurate representation of the group as well. This result is in line with the Social Sciences literature stating that emotions convey attributes such as intentions, and capabilities [22], which are also relevant for the instrumental property of cohesion and more specifically for the Task dimension [16]. Figure 6 also indicates that this model significantly improved the performances of T4 concerning the Task dimension: a significantly higher F1-score (i.e., 0.52 ±0.10 instead of 0.43 ±0.08, $p = 0.022$). Regarding the prediction of the secondary task, that is the prediction of valence, the model reached on average a weighted F1-score of 0.65 ±0.03. We can explain these performances by the fact that the models' selection, for each seed, is based on the highest weighted F1-score of the primary task (i.e., the prediction of the Social and Task dimensions). This requires a trade-off in terms of performance for the secondary task. These results indicate that the features of social interaction presented in Section III-B contain enough information to describe both phenomena. It also highlights the difficulty to predict both phenomena within the same model.
To summarize, only integrating valence following the Bottom-up approach (i.e., with the fItG_Bu model) significantly improved the performances of the fItG model for the Task dimension. This result confirms that jointly predicting the dynamics of cohesion and the valence helps to learn a shared representation of the features that brings additional information to the prediction of the Task dimension of cohesion. As stated in Vanhove and Herian's work [18], the relationships between emotions and the Task dimension exist and the model benefits from the joint training of the two tasks.

VII. CONCLUSION AND FUTURE WORK

In this paper, we presented the fItG_Bu and fItG_Td architectures to jointly learn cohesion and emotion by exploiting their interdependencies to improve the performances on the prediction of the dynamics of the Social and Task dimensions of cohesion. These 2 architectures implement the Bottom-up and Top-down approaches to group emotion. Both these architectures implement a multitask approach to jointly predict the primary tasks and the secondary task. Only the fItG_Bu architecture, however, showed significant improvements in predicting cohesion's dynamics for the Task dimension. This result implies that, in this particular setting, finding the optimal representation of an individual has more impact on the final cohesion's dynamics predictions. This model confirms the theoretical and practical benefits of multitasking as it takes full advantage of the inherent relationships between group emotion and group cohesion to improve Task cohesion's predictions. A limitation of this work is that both the architectures are specifically designed for a fixed number of persons (here 3). Adding a new group member would imply retraining the models. In the future, a computational model able to adapt itself to various sizes of groups will be developed. Furthermore, more complex labeling strategies could be conceived. Concerning cohesion, a measure of the inter-members agreement could be used to refine the labels. They could also result from the combination of self and external assessments of cohesion to minimize biases introduced by both ratings [48]. Concerning emotion, its arousal could be used to complement valence providing more fine-grained information for moving from a binary to a multiclass approach.

ACKNOWLEDGMENT

Thanks to P. Colombo, N. Lehmann-Willenbrock and A. D'Ausilio for the fruitful discussions.

REFERENCES

[1] G. A. Van Kleef, M. W. Heerdink, and A. C. Homan, "Emotional influence in groups: the dynamic nexus of affect, cognition, and behavior," *Current opinion in psychology*, vol. 17, pp. 156–161, 2017.
[2] J. S. Lerner, Y. Li, P. Valdesolo, and K. S. Kassam, "Emotion and decision making," *Annual review of psychology*, vol. 66, pp. 799–823, 2015.
[3] S. W. J. Kozlowski and D. R. Ilgen, "Enhancing the effectiveness of work groups and teams," *Psychological Science in the Public Interest*, vol. 7, no. 3, pp. 77–124, 2006.
[4] R. Grossman, S. B. Friedman, and S. Kalra, *Teamwork Processes and Emergent States*, ch. 11, pp. 243–269. John Wiley & Sons, Ltd, 2017.
[5] L. Rosh, L. R. Offermann, and R. Van Diest, "Too close for comfort? Distinguishing between team intimacy and team cohesion," *Human Resource Management Review*, vol. 22, no. 2, pp. 116–127, 2012.
[6] H. Hung and D. Gatica-Perez, "Estimating cohesion in small groups using audio-visual nonverbal behavior," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 563–575, 2010.

[7] M. C. Nanninga, Y. Zhang, N. Lehmann-Willenbrock, Z. Szlávik, and H. Hung, "Estimating verbal expressions of task and social cohesion in meetings by quantifying paralinguistic mimicry," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 206–215, Association for Computing Machinery, 2017.

[8] F. Walocha, L. Maman, M. Chetouani, and G. Varni, "Modeling dynamics of task and social cohesion from the group perspective using nonverbal motion capture-based features," in *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pp. 182–190, 2020.

[9] A. Dhall, "Emotiw 2019: Automatic emotion, engagement and cohesion prediction tasks," in *2019 International Conference on Multimodal Interaction*, pp. 546–550, 2019.

[10] B. Zou, Z. Lin, H. Wang, Y. Wang, X. Lyu, and H. Xie, "Joint prediction of group-level emotion and cohesiveness with multi-task loss," in *Proceedings of the 2020 5th International Conference on Mathematics and Artificial Intelligence*, pp. 24–28, 2020.

[11] G. Sharma, S. Ghosh, and A. Dhall, "Automatic group level affect and cohesion prediction in videos," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 161–167, IEEE, 2019.

[12] L. Festinger, S. Schachter, and K. W. Back, *Social pressures in informal groups; a study of human factors in housing*. Harper, 1950.

[13] A. V. Carron, "Cohesiveness in sport groups: Interpretations and considerations," *Journal of Sport psychology*, vol. 4, no. 2, pp. 123–138, 1982.

[14] K. Dion, "Group cohesion: From field of forces to multidimensional construct," *Group Dynamics: Theory, Research, and Practice*, vol. 4, no. 1, pp. 7–26, 2000.

[15] D. J. Beal, R. R. Cohen, M. J. Burke, and C. L. McLendon, "Cohesion and performance in groups: A meta-analytic clarification of construct relations," *Journal of Applied Psychology*, vol. 88, no. 6, pp. 989–1004, 2003.

[16] J. Severt and A. Estrada, "On the function and structure of group cohesion," in *Team Cohesion: Advances in Psychological Theory, Methods and Practice*, vol. 17, pp. 3–24, Emerald Group Publishing Limited, 2015.

[17] S. G. Barsade and D. E. Gibson, "Group emotion: A view from top and bottom.," *D. H Gruenfeld (Ed.)*, pp. 81–102, 1998.

[18] A. J. Vanhove and M. N. Herian, "Team cohesion and individual well-being: A conceptual analysis and relational framework," in *Team Cohesion: Advances in Psychological Theory, Methods and Practice*, pp. 53–82, Emerald Group Publishing Limited, 2015.

[19] E. J. Lawler, S. R. Thye, and J. Yoon, "Emotion and group cohesion in productive exchange," *American Journal of Sociology*, vol. 106, no. 3, pp. 616–657, 2000.

[20] S. G. Barsade and A. P. Knight, "Group affect," *Annual Review of Organizational Psychology and Organizational Behavior*, vol. 2, no. 1, pp. 21–46, 2015.

[21] K. Lewin, "Behavior and development as a function of the total situation," in *Manual of child psychology*, pp. 791–844, John Wiley & Sons Inc, 1946.

[22] J. C. Magee and L. Z. Tiedens, "Emotional ties that bind: The roles of valence and consistency of group emotion in inferences of cohesiveness and common fate," *Personality and Social Psychology Bulletin*, vol. 32, no. 12, pp. 1703–1715, 2006.

[23] G. Le Bon, *The crowd: A study of the popular mind*. TF Unwin, 1897.

[24] D. E. Gibson, "The struggle for reason: The sociology of emotions in organizations," *Social perspectives on emotion*, vol. 4, pp. 211–256, 1997.

[25] R. F. Bales and F. L. Strodtbeck, "Phases in group problem-solving.," *The Journal of Abnormal and Social Psychology*, vol. 46, no. 4, pp. 485–495, 1951.

[26] M. T. Braun, S. W. Kozlowski, and G. Kuljanin, "Multilevel theory, methods, and analyses in management," 2021.

[27] U. Kubasova, G. Murray, and M. Braley, "Analyzing verbal and nonverbal features for predicting group performance," in *Proc. Interspeech 2019*, pp. 1896–1900, ISCA, 2019.

[28] A. L. Gonzales, J. T. Hancock, and J. W. Pennebaker, "Language style matching as a predictor of social dynamics in small groups," *Communication Research*, vol. 37, no. 1, pp. 3–19, 2010.

[29] Y. Wang, J. Wu, J. Huang, G. Hattori, Y. Takishima, S. Wada, R. Kimura, J. Chen, and S. Kurihara, "Ldnn: Linguistic knowledge injectable deep neural network for group cohesiveness understanding," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, ICMI '20, pp. 343—-350, Association for Computing Machinery, 2020.

[30] S. Ghosh, A. Dhall, N. Sebe, and T. Gedeon, "Automatic prediction of group cohesiveness in images," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[31] D. Guo, K. Wang, J. Yang, K. Zhang, X. Peng, and Y. Qiao, "Exploring regularizations with face, body and image cues for group cohesion prediction," in *2019 International Conference on Multimodal Interaction*, ICMI '19, pp. 557–561, Association for Computing Machinery, 2019.

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[33] L. Maman, E. Ceccaldi, N. Lehmann-Willenbrock, L. Likforman-Sulem, M. Chetouani, G. Volpe, and G. Varni, "Game-on: A multimodal dataset for cohesion and group analysis," *IEEE Access*, vol. 8, pp. 124185–124203, 2020.

[34] A. V. Carron, W. N. Widmeyer, and L. R. Brawley, "The development of an instrument to assess cohesion in sport teams: The group environment questionnaire," *Journal of Sport Psychology*, vol. 7, no. 3, pp. 244–266, 1985.

[35] I. J. Roseman, "A model of appraisal in the emotion system," *Appraisal processes in emotion: Theory, methods, research*, pp. 68–91, 2001.

[36] N. Ambady and R. Rosenthal, "Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness.," *Journal of personality and social psychology*, vol. 64, no. 3, pp. 431–441, 1993.

[37] D. Gatica-Perez, L. McCowan, D. Zhang, and S. Bengio, "Detecting group interest-level in meetings," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1, pp. I/489–I/492, IEEE, 2005.

[38] E. Ceccaldi, N. Lehmann-Willenbrock, E. Volta, M. Chetouani, G. Volpe, and G. Varni, "How unitizing affects annotation of cohesion," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–7, 2019.

[39] I. Gordon, A. Gilboa, S. Cohen, N. Milstein, N. Haimovich, S. Pinhasi, and S. Siegman, "physiological and behavioral synchrony predict group cohesion and performance," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.

[40] A. Hans and E. Hans, "Kinesics, haptics and proxemics: Aspects of nonverbal communication," *IOSR Journal of Humanities and Social Science (IOSR-JHSS)*, vol. 20, no. 2, pp. 47–52, 2015.

[41] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[42] P. Müller, M. X. Huang, and A. Bulling, "Detecting low rapport during natural interactions in small groups from non-verbal behaviour," in *23rd International Conference on Intelligent User Interfaces*, pp. 153–164, 2018.

[43] A. Kasparova, O. Celiktutan, and M. Cukurova, "Inferring student engagement in collaborative problem solving from visual cues," in *Companion Publication of the 2020 International Conference on Multimodal Interaction*, ICMI '20 Companion, pp. 177–181, Association for Computing Machinery, 2020.

[44] R. G. Lord, J. F. Binning, M. C. Rush, and J. C. Thomas, "The effect of performance cues and leader behavior on questionnaire ratings of leadership behavior," *Organizational Behavior and Human Performance*, vol. 21, no. 1, pp. 27–39, 1978.

[45] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning.," in *Interspeech*, vol. 2017, pp. 1103–1107, 2017.

[46] C. Colas, O. Sigaud, and P.-Y. Oudeyer, "How many random seeds? statistical power analysis in deep reinforcement learning experiments," *arXiv preprint arXiv:1806.08295*, 2018.

[47] A. S. Yeh, "More accurate tests for the statistical significance of result differences," *CoRR*, vol. cs.CL/0008005, 2000.

[48] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.