Towards an emotional engagement model: Can affective states of a learner be automatically detected in a 1:1 learning scenario?

Nese Alyuz[1], Eda Okur[1], Ece Oktay[1], Utku Genc[1], Sinem Aslan[1], Sinem Emine Mete[1], David Stanhill[1], Bert Arnrich[2], Asli Arslan Esme[1]

[1]Intel Corporation
{nese.alyuz.civitci, eda.okur, ece.oktay, utku.genc, sinem.aslan, sinem.mete, david.stanhill, asli.arslan.esme}@intel.com

[2] Bogazici University, Turkey
bert.arnrich@boun.edu.tr

ABSTRACT

Existing Intelligent Tutoring Systems (ITSs) are unable to track affective states of learners. In this paper, we focus on the problem of emotional engagement, and propose to detect important affective states (i.e., 'Satisfied', 'Bored', and 'Confused') of a learner in real time. We collected 210 hours of data from 20 students through authentic classroom pilots. The data included information from two modalities: (1) appearance which is collected from the camera, and (2) context-performance that is derived from the content platform. In this paper, data from nine students who attended the learning sessions twice a week are analyzed. We trained separate classifiers for different modalities (appearance and context-performance), and for different types of learning sections (instructional and assessment). The results show that different sources of information are generically better representatives of engagement at different sections: For instructional sections, generic appearance classifier yields higher accuracy (55.79%); whereas context-performance classifier is more accurate for assessment sections (63.41%). Moreover, the results indicate that expression of engagement is person-specific through both of these sources, and personalized engagement models perform more accurately: When person-specific data are added to the training set, on instructional sections, 85.44% and 96.13% accuracies are achieved for appearance and context-performance, respectively. For assessment sections, the accuracies are 75.25% (appearance) and 90.24% (context-performance). When only person-specific data are employed during training, similar accuracies are achieved even with very limited data.

CCS Concepts

• **Human-centered computing→Human-computer interaction**
• **Human-centered computing→Personal computing.**

Keywords

Emotional engagement, adaptive learning, personalization, affective computing, Intelligent Tutoring Systems.

1. INTRODUCTION

Current educational systems are designed based on the needs of an industrial society [1]: "one-size-fits-all". Personalization ("accommodate-for-each") is a key to design systems capable of addressing needs of individual students in the Information Age [2]. Technology is considered as an enabler for personalization in education [3]**.** Towards this end, Intelligent Tutoring Systems (ITSs) are used to track the learning process of students by monitoring their actions, creating a learning profile for each student, and providing real-time feedback for many learning difficulties [4], [5]**.** Although such systems are capable of personalization to some extent, they lack the required empathic capabilities. We envision a novel technology - an empathic autonomous 'tutor' - playing a role similar to a 1:1 human tutor.

Relevant research indicates that engagement is positively correlated with learning [6]: The more students are engaged in learning activities, the more they learn. In [7], the overall engagement level of a student is defined as a combination of three parameters: (1) Cognitive engagement, defining the inner psychological quality during the learning process; (2) behavioral engagement, representing the learner's observable actions (e.g., OnTask/OffTask); and (3) emotional engagement, corresponding to affective states of a learner during a learning task (e.g., happiness, boredom, or confusion). In this research, we see emotional engagement detection as a trigger for personalized experience. The majority of current ITSs provide teachers with a rough idea of students' engagement in learning tasks based on interaction data between student and the content platform. However, they do not provide any concrete information about students' affective (i.e., emotional) states – especially throughout instructional tasks.

In our research, our goal is to develop the empathic autonomous 'tutor' that can closely monitor students in real-time using multiple sources of data to understand their affective states. We aim to use affective state information as a trigger for personalizing students' learning experience: For example, if a student is reading an article and identified as bored, a video representation of the content can be suggested. Or, if the student is detected as confused while solving a math question, hints can be provided for scaffolding.

The remainder of this paper is organized as follows: In Section 2, the related literature is reviewed, considering the most important challenges driving our work. The proposed methodology is outlined in Section 3, followed by a summary of our experimental results in Section 4. Section 5 highlights our conclusions and future directions for our research.

2. RELATED WORK

Although there are some efforts towards affective computing in education, some major challenges still remain unaddressed. Such challenges include the following: (1) Learning-related affective states should be considered instead of the six basic emotions [8]; (2) Data acquisition in real-life scenarios is a challenging task [9]; (3) Multimodal approaches should be employed for improved engagement modeling [10]; and (4) Model personalization is

necessary for accurate detection [11]. The remainder of this section will describe each of these challenges with literature.

2.1 Learning-related affective states

Students' affective states can influence overall learning outcomes either positively or negatively [12], [13], [14]. Recognizing and addressing such states is crucial to positively impact student learning [15], [16]. However, in a classroom where there is one teacher and many students, addressing those states for every individual in a timely manner is often unrealistic. This brings up the need for intelligent systems capable of detecting and taking actions towards students' affective states.

There is an extensive track of research on detecting facial expressions specifically focusing on basic emotions of anger, fear, sadness, happiness, disgust, and surprise as described by Ekman [17] (an exhaustive review can be found in [18]). However, a recent review of 24 studies shows that the six basic emotions are not directly applicable to learning domain [12]: Instead, affective states such as bored, confused, satisfied (i.e., delight) are commonly observed during learning [19].

There are studies focusing on the development of intelligent systems that can automatically detect students' affective states and intervene accordingly to induce positive learning outcomes [12]. For example, in [20], the binary classification problem of whether a student was interested or not during learning activity was investigated. In [21], [22], and [23], automatic recognition of frustration was investigated. In [24], students' posture is used to track boredom (low engagement) and flow (high engagement). In [25], affective states considered are confused, frustrated, engaged, bored, or neutral. In their most recent study [9], an updated list of affective states as bored, confused, frustrated, delighted, and engaged is provided. In [10], affective state detection was investigated as a four-state classification problem for detecting confidence, frustration, excitement, and interest.

We base our research on the circumplex model [26]. During the labeling phase, we used four emotional labels 'Excited', 'Calm', 'Bored', and 'Confused' corresponding to one quadrant of the circumplex model respectively. As outlined in Section 3.1.2, we merged the positive valence states of 'Excited' and 'Calm' into a single state 'Satisfied'. In Section 4.1 we provide evidence that the adapted label set increased the inter-rater agreement level.

2.2 Data acquisition in real-life scenarios

Another challenge we aim to address in this research is data acquisition in real-life scenarios. Although there has been a great interest in detecting learning-related affective states, most of these studies are limited in terms of learning scenarios and/or amount of data used. In the majority of such studies, the data collection took place in a controlled laboratory environment. The advantage of lab environments is the controlled ambience (e.g., lighting, or background) with minimal distractions [9]. However, data collected in such environments does not allow to create models that capture real complexities of classrooms. In the literature, there are only a few studies that employ in-the-wild (i.e., in a realistic classroom scenario) databases [10], [9]. Unfortunately, these databases are limited in terms of time span (i.e., 4-5 one-hour long sessions), and none are publicly available. Therefore, for our research, we collected and labeled approximately 210 hours of student data collected over 17 sessions in authentic classroom scenarios.

2.3 Multimodal approach

Environmental factors in authentic classrooms usually result in far noisier data compared to data collected in a lab, especially for appearance data. Due to distractions in real-life classroom scenarios, appearance information (i.e., face-related data) can be polluted with unusual head pose or hand gestures obstructing facial area, and thereby preventing detection based on facial features. Moreover, according to [27], identical facial configurations include significantly different emotions depending on context. It is also indicated in [28] that the interpretation of human behavioral signal requires one to know the context where it is displayed, since its interpretation is significantly context-dependent. For instance, satisfaction can be expressed as a smiling face in context of leisure time and as a neutral face with wider-eyes in context of learning, whereas a frowning action unit can be an indication of anger in context of inter-personal communication and frustrated in context of learning. Therefore contextual information from a content platform can act as a complementary source of information. However, in literature, there are only a limited number of studies including contextual information as a source of modality. In [21], the multimodal sensory information from facial expression was combined with information about a learner's activity on a computer. In [10], the use of physiological data were employed together with the contextual data from the tutoring system.

In our research, we employ contextual information coming from the content platform as the complementary modality to appearance. Here, the platform provides us information about the context (e.g. difficulty level of the content) and the performance (e.g. number of hints taken). The contextual features are explained in detail in Section 3.2.

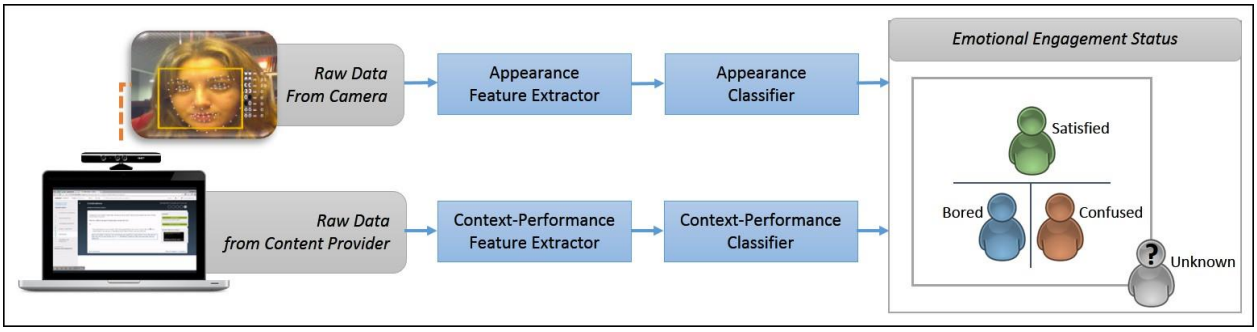


Figure 1. Overall scheme of the generic emotional engagement detector.

2.4 Model personalization

Despite of individual differences, current efforts in affect-related educational research are towards creating a generic learner-state model. However, as shown in recent research [11], [29], [30] for basic emotion recognition, personalized models can perform better in terms of detection accuracy. To address this gap in the educational research, we propose to develop a personalized emotional engagement model. Details on the proposed personalization scheme are provided in Section 3.

3. METHODOLOGY

In this paper, we aim to develop a system that can detect a learner's emotional engagement through a personalized and multi-modal approach. The system setup includes a student using a computing device (such as a PC or a tablet) equipped with a camera, and consuming educational content through a content platform. The overall scheme for the generic emotional engagement detector is given in Figure 1: The raw data acquired from camera and content platform are fed into corresponding feature extractors, and then to the classifiers. The two classification outputs define the emotional engagement status of the learner.

Once the affective state of a student is detected in an online manner, emotional engagement information can be used either directly by providing interventions for an improved learning experience of the student or by showing the teacher real-time status for each individual student through a dashboard.

3.1 Data Collection and Labeling

3.1.1 Data Collection

One of the major challenges is to collect labeled data that is necessary for model training. Towards this end, we ran authentic classroom pilots with real students from a high school in Turkey. Our target group was 9th grade students (14-15 years old). We collected data in a math course that was optionally offered as a part of this study to interested students. The lessons were scheduled twice a week with 17 sessions for 20 students in total. Among the 20 volunteering students, three of them dropped the course within the semester. Overall, nine of the students participated in the course twice a week, whereas the other eight participated once a week. At the end, around 210 hours of data were collected from these students.

Students used an online, publicly available math learning tool as a content platform in the sessions. Each of the sessions took around 60 minutes. During these sessions, the students watched instructional videos related to different math topics in the school curriculum and solved exercises (i.e., math questions) related to the topics covered. For each session, a math teacher was present in the classroom as a mentor. The specific curriculum was selected by the teacher as being appropriate for student level.

During each session, our data collection framework recorded the video of the individual students with a 3D camera (i.e., Intel® RealSense™ Camera F200), and collected the context and performance logs from the content platform. Each student worked independently in the class using a laptop computer.

3.1.2 Labeling Process

For the supervised training phase of our models and for the performance evaluation of our system, ground truth labels were necessary. The labeling was done by experts with a background in educational psychology. We incorporated Human Expert Labeling Process (HELP) as described in [31] to rigorously label student data with respect to affective states. We developed and utilized a labeling tool. The experts provide labels based on inspecting four different inputs: They simultaneously monitor individual students' videos and corresponding desktop captures; listen to audio including environment noise and students' voices; and view additional contextual information about the recording (e.g., session number, lecture topic, etc.) to decide on final labels. The experts did continuous labeling: Whenever they observe a state change in student's data, they assigned a new label.

To perform labeling, eight labelers were hired and trained by an educational researcher using HELP. For increased reliability, each recording of a student was labeled by five different labelers. The inter-rater reliability was measured after the training session and was regularly tracked during the labeling process to detect any outliers.

For the emotional engagement, we initially followed the suggestion from [8] and used four labels, each corresponding to one quadrant of the circumplex model – 'Excited', 'Calm', 'Bored' and 'Confused'. In addition to these four affective states, we also used the label 'Unknown' stating that the labeler cannot decide on the state, and the label 'N/A' (Not Available) stating that a segment is not valid either due to student is not visible or class-content is not active.

The results from the labelers' post-interviews showed that a distinction between positive and negative arousal for positive valence states was not clear. In [8] it was proposed that from an educational point of view, the two positive valence quadrants can be treated in the same way. Following this suggestion, we merged the positive valence states of 'Excited' and 'Calm' into a single state 'Satisfied'. To reinforce this decision, the inter-rater agreement between the original label set with one label per quadrant and the adapted with a single label for positive valence quadrants were compared. As the reliability coefficient for measuring the agreement among multiple raters, we utilized Krippendorff's alpha [32], and the results are given in Section 4.1.

3.1.3 Final Label Assignments

After the recorded data were labeled by the five labelers, we analyzed the labelers' decisions. Note that we used sliding windows of 8-seconds (with an overlap of 4-seconds) and treated each window as a separate instance. Hence, the labeling data with intervals defined by each labeler for each state-change, was divided into fixed instances with duration of 8-seconds. To assign a final label to each instance, majority voting is applied together with validity filtering (i.e., if there is no majority among labelers, an instance is labeled as invalid/unknown). In addition to the final labels, an agreement level for each instance is assigned to carry out further experiments with data belonging to different agreement levels. Numerical details are provided in Section 4.1.

3.2 Feature Extraction

The features used in our system refer to the segments of 8-seconds length with an overlap of 4-seconds.

3.2.1 Appearance Features

The videos of individual students were recorded with Intel® RealSense™ Camera F200 during the data collection sessions. The raw video data includes the RGB and depth streams which are used for the extraction of low-level features via Intel® RealSense™ SDK [33]: Face location and head pose in the 3D space, 2D and 3D positions of 78 facial landmarks, head pose, 22 facial expressions, and seven basic facial emotions together with sentiment values are

considered as the frame-wise features. These frame-wise features are only employed in the extraction of segment-wise higher-level features necessary for emotional engagement detection. As in [34], we extracted higher-level features including various L-estimator statistical values (e.g., tri-mean of head velocity) and energy calculations (e.g., trend of pose energy), related to head position and pose, to facial expressions, and to seven basic emotions. These robust statistical features constitute the appearance features. The groupings of appearance features used in this paper are given in Table 1.

Table 1. Appearance and context-performance feature subgroups and the corresponding feature counts.

Appearance Features	Number of Features	Examples
Tracking ratio	2	Position and pose tracking
Head position and pose	128	Trend of pose energy, median of absolute head center acceleration, standard deviation of head position, etc.
Facial expressions	32	Number of right eye raisers per segment, mean of smile, etc.
Seven basic emotions	28	Mean of anger intensity, number of joyful segments etc.
TOTAL	190	

Context-Performance Features	Number of Features	Examples
Time related	6	Time from beginning, video/attempt duration, etc.
Trial related	3	Trial number, number of trials until success, etc.
Hint related	5	Number of hints used on attempt or question, etc.
Grade related	7	Grade, correct attempt percentage, etc.
Other	3	Gender, question number from beginning, etc.
TOTAL	24	

3.2.2 Context and Performance Features

Contextual features are extracted partly from user profiles and session information we had in our database (i.e., gender, age, time of a day), in addition to the data from the content platform (i.e., video duration, exercise/trial number, time within a session). Some of these contextual features are related to the instructional sections (e.g., video duration), some are related to the assessment sections (e.g., question number), and some are related to both types of these sections (e.g., time from beginning). The performance features are extracted from the user profile data containing user characteristics provided by the content platform. Note that these performance features are all related to the assessment sections, in which the students are expected to solve exercises. These features are extracted either per each assessment section where a group of questions are solved in a row, per each question or per each attempt (i.e., each trial within a question). In general, performance features are related to the grade, the time spent, the number of trials and the number of hints taken for a question. In addition to these initial performance features, we examined features that were used in [35] with a different content platform. We adapted those features that are applicable to our platform.

Since contextual and performance features are both obtained from the content platform, we employed data fusion at feature level and concatenated these two features into one context-performance feature set. The groupings of context-performance features examined in this study are given in Table 1, together with feature counts and some exemplary features.

3.3 Uni-modal Classification

As the uni-modal classifiers, we employed Random Forest (RF) classification method [36]. The idea behind RF is that it grows many decision trees while using a randomly selected subset of training data for each tree. Moreover, a randomly selected subset of features is used to split each node. The final class for a test sample is assigned by the majority vote among all trees. The advantages of using RF as a classifier are that there is no need for pruning and cross validation, and over-fitting is not an issue. For all these reasons, Random Forest (RF) with 100 classification trees is selected to be the final classification method. For the two different modalities, we trained two separate RF classifiers: (1) Appearance classifier, and (2) context-performance classifier.

3.4 Model Personalization

As empirically shown in [29], [30], person-specific models achieve significant improvement over person-independent models if the subject-specific data are sufficient for model training. For our future work, we envision to obtain personalized models for emotional engagement detection through online self-labeling of the person-specific data: During the lecture, the individual students will be asked through intervention pop-ups to self-label themselves at randomized time points. The self-labeling interface will also be embedded into the content platform and will be permanently reachable, so that the students can give labels any time. At the end of each session, the training data will be updated with the labeled person-specific data and the model will be retrained.

In this paper, we investigated the improvement that could be achieved by personalized models: Since self-labels are not available for the current dataset, we considered the ground truth labels as self-labels. We applied personalization through including person-specific data during the training phase, for both of the uni-modal classifiers. We experimented with two different approaches on how to include the person-specific data: (1) 'Adapted', and (2) 'Personal'. In 'Adapted', we augmented the initial training set of the 'Generic' model, collected from a different set of students, with the acquired and labeled instances of the test subject. In 'Personal', person-specific training sets are generated by using only the personal data. The aim of the 'Adapted' model is to merge the capabilities of the 'Generic' detector (which is trained on a large database) with the characteristics residing in the person-specific data. However, if the personal data is sufficient for training, the 'Personal' model would be better to represent person-specific behaviors. In Figure 2, the training sets used in different models are visualized. In Section 4.3, the preliminary experiments to show the need for personalized models are summarized.

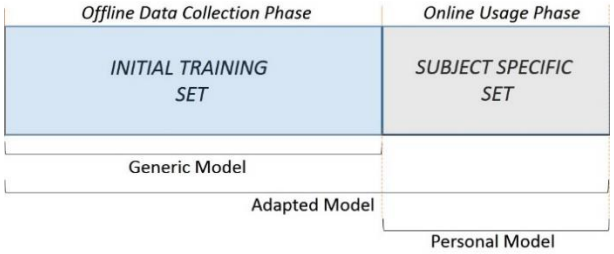
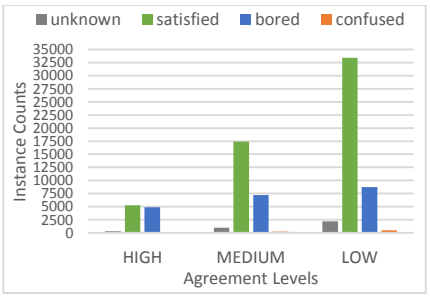
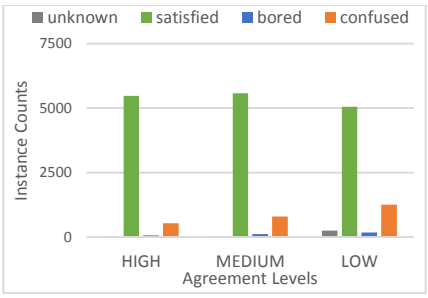


Figure 2. Training sets used for different models: (1) Generic, (2) Adapted, and (3) Personal.



(a) Instructional



(b) Assessment

Figure 3. Distribution of samples for different agreement levels of (1) High (5/5), (2) Medium (4/5), and (3) Low (3/5), for (a) Instructional, and for (b) Assessment sections.

4. EXPERIMENTAL RESULTS

For emotional engagement detection, we experimented with two different modalities: (1) Appearance and (2) context-performance. The learning content included two different types of sections: (1) Instructional, and (2) assessment. Currently, these two section types are considered as two separate problems, and two separate models are constructed. For each different modality and each section type, we experimented with three different models: (1) 'Generic', where the training set contains data of students separate from the test subject in a leave-one-out manner; (2) 'Adapted', where the training set is augmented with the subject-specific data; and (3) 'Personal', where subject-specific training sets are constructed using only the subject-specific data.

4.1 Labeler Agreement Analysis

As outlined in Section 3.1.2, during the labeling phase, the four emotional labels, 'Excited', 'Calm', 'Bored', 'Confused', and two other labels 'Unknown', 'N/A' were used (i.e., 4+2 states). As stated in the labelers' post-interviews, the distinction between high and low arousal for positive valence states was not always clear. This experience was in line with the proposal of [8] to treat positive valence quadrants in the same way. Following this suggestion, we merged the two positive valence states 'Excited' and 'Calm' into a single state 'Satisfied' (i.e., 3+2 states). To reinforce our decision, we compared the inter-rater agreement between the original 4+2 label set and the adapted 3+2 one: ten hours of data collected from four subjects were labeled by three persons in both ways. The inter-rater agreement level according to Krippendorff's alpha [32] was 0.2 in the original 4+2 label set and it increased to 0.4 in the adapted 3+2 label set.

Since multiple labelers were employed for the labeling process, it was necessary to assign a final label to each of the instances. For this process, we applied majority voting. However, since emotional labeling is a subjective task and the inter-rater agreement level for the small experimentation outlined above is not sufficiently high (below 0.8), we conducted a filtering over the traditional majority voting: We computed the ratio of the agreement, and grouped instances accordingly. We had three agreement levels: (1) High, (2) medium, and (3) low; for 5/5, 4/5, and 3/5 majority votes, respectively. The other samples were regarded as instances of disagreement. The data distribution of agreement levels are visualized in Figure 3, in (a) for instructional, and in (b) for assessment samples. As can be seen in these figures, the number of 'Confused' samples for the instructional, and the number of 'Bored' samples for the assessment sections were too few. Therefore, we have discarded 'Confused' class for the instruction sections, and 'Bored' for the assessment sections. Furthermore, we decided to use High-Medium agreement level for the instructional sections. However, for the assessment, we used all of the agreement levels, since the assessment sections were short in general and this would have led to limited number of samples in total. In addition to agreement samples, we included the disagreement instances as representatives of the 'Unknown' class.

To investigate how the performance of the personalized models changed, we selected students who attended most of the sessions (i.e., twice a week). Therefore, in the experiments summarized in this paper, the data from nine of the students are utilized.

4.2 Generic Classification Results

Although the main problem addressed in this paper is model personalization, we also included results on the generic model for comparative purposes. For each section type (instructional vs. assessment) and for each modality (appearance vs. context-performance), separate RF classifiers are trained. The available data of each student are divided into training and test sample sets, as approximately 80% and 20% of the whole data, respectively. For each individual, we carried out leave-one-subject-out approach, where the training samples of all the other students are utilized to construct the training set of that individual's classifiers. Due to data imbalance, we also experimented with 10-fold random down-sampling to construct balanced training sets: For each student, the instance count of the limiting class (with the minimum number of training samples) is used for random instance selection per class, and the random selection is carried out for ten times. As results, we reported F1 measure which incorporates both precision and recall values.

4.3 Personalization Results

In this paper, we employed the ground truth labels to construct person-specific labeled sets, necessary for the personalization experiments. When constructing person-specific models, we considered two approaches: (1) 'Adapted', where the training set of the generic model is augmented with the person-specific data; (2) 'Personal', where only the person-specific data is used in the training phase. The results for the three models of 'Generic', 'Adapted', and 'Personal' are compared in Table 2 and Table 3, for instructional and assessment sections, respectively: The average numbers of test instances for each student are given in column 2. The average number of instances used in training, and average F1 values for the appearance and the context-performance classifiers are given in columns 3-5, in columns 6-8, and in columns 9-11; for 'Generic', 'Adapted', and 'Personal' models, respectively. As the

Table 2. Engagement detection results (F1-measures) for *instructional* sections on Appearance (Appr.) and context-performance (C-P) modalities, using: (1) the 'Generic', (2) the 'Adapted', and (3) the 'Personal' models.

Classes	Average Test Size	GENERIC MODEL			ADAPTED MODEL			PERSONAL MODEL		
		Average Training Size	Appr. (%-F1)	C-P (%-F1)	Average Training Size	Appr. (%-F1)	C-P (%-F1)	Average Training Size	Appr. (%-F1)	C-P (%-F1)
Unknown	12	967	10.73	9.62	1018	24.85	72.97	51	33.04	85.38
Satisfied	336	967	61.04	55.76	2273	87.63	96.12	1305	89.65	97.18
Bored	151	967	44.93	39.68	1542	70.91	93.33	575	73.54	94.41
OVERALL	499	2901	55.79	49.50	4833	85.44	96.13	1931	89.30	97.32

Table 3. Engagement detection results (F1-measures) for *assessment* sessions on Appearance (Appr.) and context-performance (C-P) modalities, using: (1) the 'Generic', (2) the 'Adapted', and (3) the 'Personal' models.

Classes	Average Test Size	GENERIC MODEL			ADAPTED MODEL			PERSONAL MODEL		
		Average Training Size	Appr. (%-F1)	C-P (%-F1)	Average Training Size	Appr. (%-F1)	C-P (%-F1)	Average Training Size	Appr. (%-F1)	C-P (%-F1)
Unknown	88	1886	33.53	27.94	2211	47.21	72.02	324	49.48	72.75
Satisfied	264	1886	60.58	76.32	2884	83.43	94.04	997	83.79	94.39
Confused	43	1886	17.12	46.59	2044	37.64	82.05	158	44.04	85.01
OVERALL	395	5658	48.12	63.41	7139	75.25	90.24	1479	76.37	90.89

overall results (last rows) in Table 2 and Table 3 indicate, it is better to include person-specific data in the training set ('Adapted'), and it is much better to obtain fully-personal models ('Personal'). Therefore, it can be stated that the information residing both in the appearance and in the context-performance modalities is specific for each person. When the appearance and context-performance classifiers are compared, the results show that the context-performance classifiers' improvement is more significant, and the context-performance features can achieve better personal models with equal amount of data. On the other hand, the appearance classifiers need more subject-specific data to achieve similar accuracies to the context-performance modality. The need for more personal data is evident especially for the 'Unknown' class (for both modalities and for both section types). In addition, for assessment sections, the number of 'Confused' samples is a limiting factor for the 'Personal' appearance classifier (Table 3, row 4).

5. CONCLUSIONS AND FUTURE WORK

The aim of this work is to detect emotional engagement of a student while the learner is consuming educational content. In this paper, we investigated appearance and context-performance modalities. For a better understanding of the classification performance, we treated instructional and assessment sections separately. For the different modalities and different section types, we experimented with three models: (1) 'Generic', (2) 'Adapted', and (3) 'Personal'. The results of the generic models showed us that appearance is more informative for the instructional sections (55.79% vs. 49.50%), whereas with the presence of performance-related features for the assessment sections, context-performance modality becomes more representative (63.41% vs. 48.12%). As the personalization experiments indicated, information included in both of the modalities are person-specific, thus model personalization is a must to obtain highly performing emotional engagement models. Context-performance classifiers achieve high improvement even with limited personal data (90.89-97.32%), whereas improvement for the appearance modality is lower (76.37-89.30%) and requires more person-specific instances to achieve accuracies as high as context-performance. For both modalities, 'Generic' models can be used for emotional engagement detection, if no person-specific data are available. Through acquisition of personal data, however, 'Adapted' models should be utilized. After sufficient amount of person-specific data are collected, 'Personal' models should be preferred for improved accuracies.

In future work, we will investigate fusion strategies to merge appearance and context-performance modalities. To further increase the dataset volume and to validate personalization experiments with real self-labels, we are currently designing a new data collection pilot. We are redesigning course content so that data imbalance can be decreased while increasing samples for 'Confused' and 'Bored' classes. Moreover, we are investigating strategies for a better representation of the 'Unknown' class. We are planning on including bio-sensors as an additional modality. We are working on strategies for personalization, where the need for manual or self-labeling is minimized. Moreover, we are working on generic and personalized feature selection methods to identify important traits.

6. REFERENCES

[1] W. R. Watson, S. L. Watson and C. M. Reigeluth, "Education 3.0: Breaking the mold with technology," *Interactive Learning Environments*, vol. 23, no. 3, pp. 332-343, 2015.

[2] S. Aslan and C. M. Reigeluth, " A trip to the past and future of educational computing: Understanding its evolution," *Contemporary Educational Technology*, vol. 2, no. 1, pp. 1-17, 2011.

[3] M. Martinez, "Designing learning objects to personalize learning," *The instructional use of learning objects*, pp. 151-171, 2002.

[4] G. Paviotti, P. G. Rossi and D. Zarka, Intelligent tutoring systems: an overview, Pensa Multimedia, 2012.

[5] F. F. Burton, Foundations of Intelligent Tutoring Systems, 2013.

[6] R. M. Carini, G. D. Kuh and S. P. Klein, "Student engagement and student learning: testing the linkages," *Research in Higher Education*, vol. 47, no. 1, pp. 1-32, 2006.

[7] J. A. Fredricks, P. C. Blumenfeld and A. H. Paris, "School engagement: Potential of the concept, state of the evidence," *Review of Educational Research*, vol. 74, no. 1, pp. 59-109, 2004.

[8] B. Woolf, W. Burleson, I. Arroyo, T. Dragon, D. Cooper and R. Picard, "Affect-aware tutors: recognising and responding to student affect," *Int. Journal of Learning Technology*, vol. 4, no. 3, pp. 129-164, 2009.

[9] N. Bosch, S. D'Mello, R. Baker, J. Ocumpaugh, V. Shute, M. Ventura and W. Zhao, "Automatic detection of learning-centered afective states in the wild," in *Int. Conf. on Intelligent User Interfaces*, 2015.

[10] I. Arroyo, D. G. Cooper, W. Burleson, B. P. Woolf, K. Muldner and R. Christopherson, "Emotion sensors go to school," *Artificial Intelligence in Education (AIED)*, vol. 200, pp. 17-24, 2009.

[11] J. Chen, X. Kiu, P. Tu and A. Aragones, "Learning person-specific models for facial expression and action unit recognition," *Pattern Recognition Letters*, vol. 34, 2013.

[12] S. D'Mello, "A selective meta-analysis on the relative incidence of discrete affetcive states during learning with technology," *Journal of Educational Psychology*, vol. 105, no. 4, p. 1082, 2013.

[13] A. C. Frenzel, R. Pekrun and T. Goetz, "Perceived learning environment and students' emotional experiences: a multilevel analysis of mathematics classrooms," *Learning and Instruction*, vol. 17, no. 5, pp. 478-493, 2007.

[14] P. Schutz and R. Pekrun, Emotion in Education, Academic Press, 2007.

[15] K. D. Sidney, S. D. Craig, B. Gholson, S. Franklin, R. W. Picard and A. C. Graesser, "Integrating affect sensors in an intelligent tutoring system," in *Affetive Interactions: The Computer in the Affective Loop Workshop*, 2005.

[16] D. Goleman, Emotional Intelligence, New York: Bantam Books, 1995.

[17] P. Ekman and E. L. Rosenberg, What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS), Oxford University Press, 1997.

[18] R. Calvo and S. D'Mello, "Affect detection: an interdisciplinary review of models, methods, and their applications," *Trans. on Affective Computing*, vol. 1, no. 1, pp. 18-37, 2010.

[19] J. Ocumpaugh, "Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual," 2015.

[20] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in *Int. Conf. on Multimedia*, 2005.

[21] A. Kapoor, W. Burleson and R. W. Picard, "Automatic prediction of frustration," *Int. Journal of Human-Computer Studies*, vol. 65, no. 8, pp. 724-736, 2007.

[22] M. E. Hoque, D. J. McDuff and R. W. Picard, "Exploring temporal patterns in classifying frustrated and delighted smiles," *Trans. on Affective Computing*, vol. 65, no. 8, pp. 323-334, 2012.

[23] J. F. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe and J. C. Lester, "Automatically recognizing facial indicators of frustration: a learning-centric analysis," in *Affective Computing and Intelligent Interaction*, 2013.

[24] S. D'Mello, P. Chipman and A. Graesser, "Posture as a predictor of learner's affective engagement," in *Annual Cognitive Science Society*, 2007.

[25] N. Bosch, Y. Chen and S. D'Mello, "It's written on your face: detection affective states from facial expressions while learning computer programming," *Intelligent Tutoring Systems*, pp. 39-44, 2014.

[26] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, p. 1161, 1980.

[27] H. Aviezer, R. R. Hassin, J. Ryan, C. Grady, J. Susskind, A. Anderson and S. Bentin, "Angry, disgusted, or afraid? Studies on the malleability of emotion perception," *Psychological Science*, vol. 19, no. 7, pp. 724-732, 2008.

[28] Z. Zeng, M. Pantic, G. Roisman and T. S. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions," *Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39-58, 2009.

[29] I. Cohen, N. Sebe, A. Garg, L. S. Chen and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision and Image Understanding*, vol. 91, no. 1, pp. 160-187, 2003.

[30] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic and K. Scherer, "The first facial expression recognition and analysis challenge," in *Int. Conf. on Automatic Face and Gesture Recognition and Workshops*, 2011.

[31] S. Aslan, S. E. Mete, E. Okur, E. Oktay, N. Alyuz, U. Genc, D. Stanhill and A. A. Esme, "Human Expert Labeling Process (HELP): Towards a reliable higher-order user state labeling by human experts," in *Int. Conf. on Intelligent Tutoring Systems (ITS) - Workshops*, 2016.

[32] K. Krippendorff, "On the reliability of unitizing continuous data," *Sociological Methodology*, pp. 47-76, 1995.

[33] Intel Corporation, "Intel RealSense SDK: Design Guidelines," 2014. [Online]. Available: https://software.intel.com/sites/default/files/managed/27/50/Intel%20RealSense%20SDK%20Design%20Guidelines%20F200%20v2.pdf.

[34] L. Nanxiang and C. Busso, "Analysis of facial features of drivers under cognitive and visual distractions," in *Int. Conf. on Multimedia and Expo (ICME)*, 2013.

[35] Z. A. Pardos, R. S. Baker, M. O. San Pedro, S. M. Gowda and S. M. Gowda, "Affective states and state tests: investigating how affect throughout the school year predicts end of year learning outcomes," in *Int. Conf. on Learning Analytics and Knowledge*, 2013.

[36] C. Chen, A. Liaw and L. Breiman, Using random forest to learn imbalanced data, University of California, Berkeley, 2004.