Affect-driven Learning Outcomes Prediction in Intelligent Tutoring Systems

Ajjen Joshi[1], Danielle Allessio[2], John Magee[3], Jacob Whitehill[4], Ivon Arroyo[4], Beverly Woolf[2], Stan Sclaroff[1], Margrit Betke[1]

[1] Department of Computer Science, Boston University, Boston, MA, USA

[2] College of Information and Computer Sciences, University of Massachussets, Amherst, USA

[3] Department of Mathematics and Computer Science, Clark University, Worcester, MA, USA

[4] Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA, USA

***Abstract*— Equipping an Intelligent Tutoring System (ITS) with the ability to interpret affective signals from students could potentially improve the learning experience of students by enabling the tutor to monitor the students' progress and provide timely interventions as well as present appropriate affective reactions via a virtual tutor. Most ITSs equipped with affect modeling capabilities attempt to predict the emotional state of users. However, the focus in this work is instead on trying to directly predict the learning outcomes of students from a stream of video capturing the students faces as they work on a set of math problems. Using facial features extracted from a video stream, we train classifiers to directly predict the success or failure of a student's attempt to answer a question while the student has just begun to work on the problem. In this work, we first introduce a novel dataset of student interactions with MathSpring, a popular ITS. We provide an exploratory analysis of the different problem outcome classes using typical facial action unit activations. We develop baseline models to predict the problem outcome labels of students solving math problems and discuss how early problem outcome labels can be forecasted and utilized to provide possible interventions.**

I. INTRODUCTION

An interesting research question for automated affect analysis in the education domain is to inquire about how modeling student affect during digital learning experiences can be utilized to positively impact the student's overall learning experience. Intelligent tutoring systems (ITSs) have been developed with the aim of providing individualized learning experiences to users. One of the goals of an ITS is to build models of the student engaged in learning and the ITS adapting its support mechanisms to personalize the teaching [13]. Students experience a variety of emotions, such as interest, surprise, boredom, frustration, confusion and anxiety, during learning [8] and these displayed emotions correlate well with their achievement in the learning task [17]. Equipping an ITS with the ability to interpret such affective signals could potentially enable it to monitor the students' progress, provide timely interventions and present appropriate affective reactions via a virtual tutor. For example, machine learning classifiers can be trained to recognize the subtle differences in facial behavior when a student requires hints to solve a problem (Figure 1), so that the ITS can intervene accordingly.

Affective tutoring systems (ATSs) are ITSs that use one or more sensors to observe the student in order to infer his or her emotional state while using the ATS. For example, ATSs such as EER-Tutor [20] and FERMAT [21] automatically recognize basic emotions, such as happiness, anger and disgust. In addition to the basic emotions, some ITSs (e.g. AutoTutor [8], Guru Tutor [16]) can classify learning-specific emotional states, such as engagement, concentration, confusion, boredom and frustration. Vision-based sensors such as webcams are suitable for capturing the facial dynamics of students as they are readily available in the most common platforms used for interacting with ITSs (e.g. phones, tablets, laptops) and are less invasive than other sensors, such as wearable devices that measure physiological signals like skin conductivity, heart rate, muscle activity or pressure-sensitive chairs that measure posture.

Fig. 1. Example images of student when she solved problem on first attempt (SOF) (top row), and when she required hints to solve the problem (SHINT) (bottom row).

The availability of domain-specific education datasets that can be shared by researchers to develop, improve, and evaluate affect-sensing machine learning algorithms can accelerate development of effective affect analysis algorithms. Considering the dearth of large-scale, publicly available affect datasets in learning and education settings, we first collected a facial affect dataset of videos of students interacting with MathSpring [1], a web-based mathematics ITS. Most ATSs equipped with affect modeling capabilities attempt to predict the emotional state of users. However, our focus is instead on trying to directly predict the learning outcomes of students. That is, using facial features extracted from a video stream, we train classifiers to directly predict the success or failure of a student's attempt to answer a question.

In this paper, we describe how and why we collected a

978-1-7281-0089-0/19/$31.00

new dataset of student interactions with MathSpring. We processed this raw data into a supervised machine learning dataset, where each data instance corresponds to a videoclip of a student working on a single problem and its corresponding label is the student's problem solving behavior (e.g. ask for hints, solves the problem). We provide an exploratory affective analysis of the different problem outcome classes using average facial action unit activations and discuss a few observed trends. Finally, we present baseline models to investigate the problem of trying to directly predict the learning outcome of students solely from affect signals.

Our contributions in this paper are twofold. First, we introduce a unique and novel dataset consisting of students videos of 1596 interactions, extracted from more than 30 hours of raw video data. We wish to make this dataset publicly accessible in order to encourage and foster research in the intersection of the education and computer vision communities. We also provide a set of baseline results of predicting student learning outcomes solely from facial affect signals and provide a preliminary analysis of the data, discussing potential proactive interventions such affect-sensitive models would enable.

II. RELATED WORK

One goal of ITSs is to provide a platform capable of delivering a personalized learning experience as per the requirements of the student [13]. A popular example of an ITS is MathSpring [1], formerly known as Wayang Outpost, which is a web-based ITS for learning mathematics concepts for middle and high school students.

An important source of information upon which to provide personalized feedback is the student's affective state. Students display a variety of emotions, such as interest, flow, surprise, anger, boredom, frustration, confusion and anxiety, during learning [8]. Emotions felt and displayed by students correlate well with their achievement in the learning task [17]. In recent years, advances in computer vision and machine learning have led to the development of fast and robust facial expression analysis tools [3]. Adapting to student affective states as measured by ITSs has been shown to improve their effectiveness [4], [7]. Grafsgaard et al. [12] showed that different facial expressions correspond to different learning experiences.

Student affect has been modeled using a variety of signals including student self-reports [19] and log data [5]. A common signal channel used by ITSs to model affective states is camera-based captures of students' faces. For example, EER-Tutor [20] and FERMAT tutor [21] track the facial features of the users with a video camera to classify the user's face to basic emotional states such as happy, smiling, angry and neutral. In contrast to the well-studied basic emotions, it has been shown that learning-centric emotions such as boredom, engagement and confusion feature more prominently during the process of interacting with ITSs [2], [10]. AutoTutor [8] uses a video camera as well as a pressure-sensitive chair to recognize learning-centric emotional states such as flow, confusion, boredom, frustration and eureka. Guru Tutor [16] utilizes a video camera and eye tracker to measure a student's level of interest and boredom. Whitehill et al. [18] presented computer-vision based techniques for automatic engagement detection. D'Mello et al. [9] introduced an advanced, analytic and automated approach to measure engagement at fine-grained temporal resolutions.

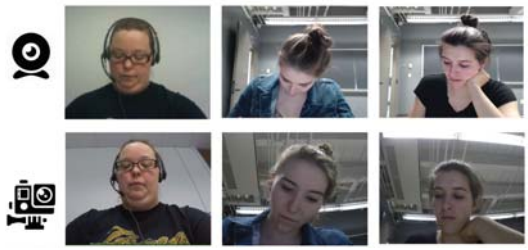

Fig. 2. Example images of dataset collected from different modalities: A front facing webcam captures the subject while she looks at the screen (top row), a secondary GoPro camera is placed at an angle on the laptop trackpad in order to capture the student's face when she faces down to work on the problem at hand (bottom row).

Although research in automated affect analysis has a long history, the few publicly available datasets in the education and learning domain are often limited to user engagement prediction [6], [15]. Instead of using affect signals to predict the user's emotional state, we wish to directly forecast student learning outcomes, which makes our dataset unique.

III. DATASET COLLECTION AND ANNOTATION

The dataset consists of video recordings of college students participating in problem-solving sessions in MathSpring, a popular browser based ITS intended to aid students in the learning of mathematics concepts. A total of 30 undergraduate college students (4 males, 26 females) participated in the study, with several students taking part in multiple sessions, each of which lasted approximately one hour. In total, 38 student sessions were recorded, from which 1596 problem samples were extracted.

The data was collected in a classroom setting, where the students were asked to solve MathSpring problems on a laptop, while being recorded by two cameras: the laptop webcam along with an auxiliary GoPro camera placed on the trackpad of the laptop (Figure 2). The webcam has a good view of the participant's face while they read problems and interact with the ITS, but usually loses view of the face when students look down to work on problems on paper on the desk. The secondary camera captures the faces of participants during the time while they are looking down and their faces are not visible to the webcam. All participants provided consent for their recordings to be included in a public dataset. The protocol was approved by IRBs at the participating institutions. In addition, mouse location trajectories and clicks were also captured and logged by the MathSpring interface. A third video stream captured the activity on the screen, including the user interface of the ITS and the user's mouse interactions with it.

Each data instance consists of a video clip of the student working on a single problem. These were obtained by

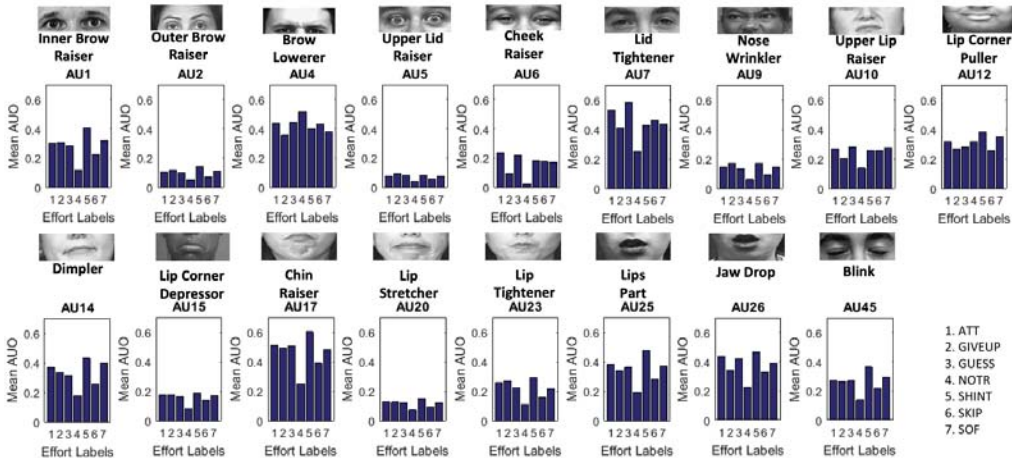

Fig. 3. Average Action Unit Occurrence distributed according to effort labels. For each subplot, the x-axis represents the 7 student learning outcome classes whereas the y-axis represents the mean AUO score as defined in Eq. 11. Grayscale images depicting the AUs were obtained from https://www.cs.cmu.edu/?face/facs.htm

TABLE I

THE NUMBER OF OCCURRENCES AND AVERAGE TIME TAKEN TO COMPLETE FOR EACH CLASS IN THE DATASET.

Class	Frequency	Average Time
ATT	172	38 seconds
GIVEUP	13	64 seconds
GUESS	148	39 seconds
NOTR	81	10 seconds
SHINT	159	87 seconds
SKIP	99	18 seconds
SOF	919	28 seconds

trimming the raw videos based on problem start and end times recorded in MathSpring's log file. The class label for each data instance is the problem-solving effort outcome label, which is automatically generated by MathSpring. The labels are: ATT (student did not see any hints but solved the question after 1 incorrect attempt), GIVEUP (student performed some action but did not solve the problem at all), GUESS (student did not see hints, but solved the question after greater than 1 incorrect attempts), NOTR (student performed some action, but the first action was too rapid for him to have read the problem), SHINT (student eventually got the correct answer after seeing one or more hints), SKIP (student skipped problem with no action) and SOF (student answered correctly in first attempt, without seeing any hints).

IV. EXPLORATORY DATA ANALYSIS

We visualized how often and with what intensity various Facial Action Units (AUs) occurred on average for the different effort classes of the entire dataset. For each data instance, we aggregated AU presence values weighted by their respective intensity values and normalized them by the total number of frames in which the face was detected:

$$AUO_a = \frac{1}{N} \sum_j AUI_a^j \times AUP_a^j, \tag{1}$$

where, AUO_a represents the mean Action Unit Occurence for AU a of the video, N represents the number of frames in the video in which the face was detected and AUI_a^j and AUP_a^j represent the presence and intensity values of AU_a for frame j respectively.

For all videos, we computed the mean AUO for 17 AUs whose presence and intensities were detected by OpenFace [3] and plotted them separated by effort classes (Figure 3). From the mean AUO plots, we can observe some interesting trends. AUs 4 (Brow lowerer), 7 (Lid tightener) and 17 (Chin raiser) are activated comparatively highly across all effort labels, whereas AUs 2 (outer brow raiser), 5 (upper lid raiser) and 20 (lip stretcher) are not. It is interesting to note that AUs 2, 5 and 20 are associated with the emotions of fear and surprise. Moreover, the mean AUOs across all AUs tend to be higher for instances labeled SHINT compared to inputs labeled SOF, indicating that participants display more affective expressiveness when requiring hints to solve a problem compared to when they solve them at first attempt.

V. BASELINE MODELS

The input to our baseline models consists of variable-length webcam videoclips of participants working on MathSpring problems. A significant proportion of the frames contain full frontal faces of the subject, representing times when they are interacting with the ITS (i.e., reading the problem, thinking about the solution, answering the question and interacting with the on-screen educational avatar). For the baseline models, frames where the frontal face of the subject could not be detected due to occlusion by the hand or severe out-of-plane rotation of the head were discarded from the training and testing procedures.

Feature Representation: For each frame of all the videos in the dataset, 18 AU presence and 17 AU intensity values, along with head-pose and eye-gaze vectors, are extracted using OpenFace [3]. In order to compute an aggregate feature representation, we used statistics (mean, standard deviation, min and max) for each feature as well as statistics for their derivatives to produce a uniform length 376-dimensional

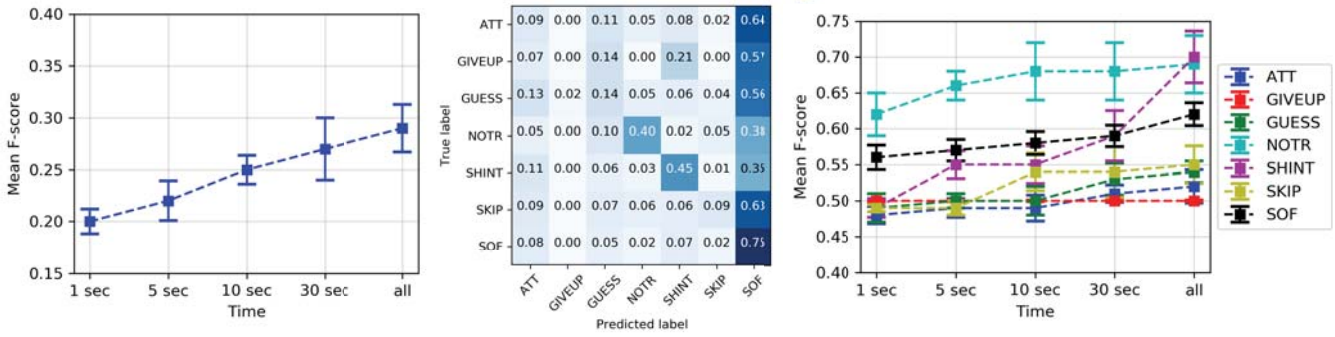

Fig. 4. Mean F-scores for a multiclass classifier (Left), Confusion matrix for the 7-class student effort prediction (Middle), and mean F-scores for one-vs-all binary classifiers trained for different problem outcome labels (Right).

feature representation. The derivatives capture the change in feature activations at each timestep. The mean, standard deviation, min and max are representative summary statistics of the variable-length features and capture the distribution and range of values for each feature, which can be used by the classifier to distinguish samples from different classes.

Experiments: Due to class imbalance, we trained and tested our models on 5 random, stratified 75/25 splits of the data for all our experiments. We trained a multi-layer perceptron with 2 hidden layers, each with 100 activation nodes, using Adam [14].

Ideally, an affect-sensitive model should be able to accurately predict the effort label of the user as early as possible, in order to enable quick and effective interventions by the ITS. Therefore, we tested our classification models when only a fraction of the data is observed during test time. In order to do so, we first trained models on the first 1, 5, 10 and 30 seconds, as well as the entire length of the input and tested them on corresponding test conditions.

We first trained a model for the multi-class effort prediction task. Our baseline model trained and tested on the entire length of the input obtained a mean accuracy of 0.54 and a mean F-score of 0.27. Given that our dataset is imbalanced with more than half the samples corresponding to the 'SOF' label, the predictions of our model are heavily biased towards that label, as is evident in the figure depicting the normalized confusion matrix (Figure 4 Middle).

We also trained individual one-vs-all binary classifiers for all effort labels. A model capable of predicting each of these indicators can help the ITS make decisions with regard to providing proactive interventions. For example, a model that can successfully predict whether a student can answer the question on the first attempt can prompt the ITS to increase difficulty levels of subsequent questions. Similarly, if the student is predicted to require hints to solve the problem, the ITS can proactively offer a hint before the student asks for it. We find that baseline binary models perform better when predicting SHINT, NOTR and SOF, indicating that facial affect signals displayed during problem-solving corresponding to these labels are the most discriminative.

In Figures 4 Left and 4 Right, we plotted the F-scores for the various problem outcome labels, as obtained by the models, when predicting problem outcomes after observing 1 second, 5 seconds, 10 seconds, 30 seconds as well as the entire length of the data instance, for the multiclass and binary classification settings respectively. We can observe that model performance, expectedly, increases as features computed from longer temporal sequences are available. Based on these experiments, our baseline models are better at accurately predicting SHINT, NOTR and SOF compared to predicting ATT, GIVEUP, GUESS and SKIP.

VI. CONCLUSIONS AND FUTURE WORKS

We investigated the problem of trying to predict the learning outcome of students from facial affect signals, based on a novel dataset of videos of students interacting with MathSpring, a popular web-based ITS. We developed models to directly predict the learning behavior of students from concise action unit-based feature representations that capture the facial affect dynamics of the input video.

There are several avenues for improving performance obtained by the baseline models. A multi-modal model that utilizes signals from all streams of information in the dataset including the GoPro video stream, the mouse movements and clicks, as well as the video stream of the screen activity will probably result in better predictive performance. Moreover, training models that explicitly utilize the temporal dynamics of how facial behavior evolves over the duration of the student's interaction with the ITS could potentially yield further improvements in model performance. Finally, the biggest challenge in ATSs is to utilize these affect-sensitive models to provide appropriate and effective interventions that quantifiably improve the learning experience. There have been some recent works that have ventured in this direction [11]. In future work, we plan to provide personalized interventions in MathSpring based on the proposed affect analysis models, and conduct experiments to validate the effectiveness of the interventions. Lessons learnt from this initial analysis will also inform future data collection strategies. We intend to use richer data sets to investigate whether the system can predict changes in student learning behaviors and strategies.

VII. ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their constructive feedback. This work was supported in part by National Science Foundation grant 1551572.

REFERENCES

[1] Ivon Arroyo, Carole Beal, Tom Murray, Rena Walles, and Beverly P Woolf. Web-based intelligent multimedia tutoring for high stakes achievement tests. In *International Conference on Intelligent Tutoring Systems*, pages 468–477. Springer, 2004.

[2] Ryan S Baker, Sidney K D'Mello, Ma Mercedes T Rodrigo, and Arthur C Graesser. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4):223–241, 2010.

[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. OpenFace: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE, 2016.

[4] Min Chi, Kurt VanLehn, Diane Litman, and Pamela Jordan. Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Modeling and User-Adapted Interaction*, 21(1-2):137–180, 2011.

[5] Seth Corrigan, Tiffany Barkley, and Zachary Pardos. Dynamic approaches to modeling student affect and its changing role in learning and performance. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 92–103. Springer, 2015.

[6] Arjun D'Cunha, Abhay Gupta, Kamal Awasthi, and Vineeth Balasubramanian. Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885*, 2016.

[7] Matt Dennis, Judith Masthoff, and Chris Mellish. Adapting performance feedback to a learner's conscientiousness. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 297–302. Springer, 2012.

[8] S D'Mello, A Graesser, and RW Picard. Toward an affect-sensitive AutoTutor. *IEEE Intelligent Systems*, 4(22):53–61, 2007.

[9] Sidney D'Mello, Ed Dieterle, and Angela Duckworth. Advanced, analytic, automated (aaa) measurement of engagement during learning. *Educational Psychologist*, 52(2):104–123, 2017.

[10] Sidney D'Mello, Blair Lehman, Reinhard Pekrun, and Art Graesser. Confusion can be beneficial for learning. *Learning and Instruction*, 29:153–170, 2014.

[11] Goren Gordon, Samuel Spaulding, Jacqueline Kory Westlund, Jin Joo Lee, Luke Plummer, Marayna Martinez, Madhurima Das, and Cynthia Breazeal. Affective personalization of a social robot tutor for children's second language skills. In *AAAI*, pages 3951–3957, 2016.

[12] Joseph Grafsgaard, Joseph B Wiggins, Kristy Elizabeth Boyer, Eric N Wiebe, and James Lester. Automatically recognizing facial expression: Predicting engagement and frustration. In *Educational Data Mining 2013*, 2013.

[13] Danita Hartley and Antonija Mitrovic. Supporting learning by opening the student model. In *International Conference on Intelligent Tutoring Systems*, pages 453–462. Springer, 2002.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] Aamir Mustafa, Amanjot Kaur, Love Mehta, and Abhinav Dhall. Prediction and localization of student engagement in the wild. *arXiv preprint arXiv:1804.00858*, 2018.

[16] Andrew M Olney, Sidney D'Mello, Natalie Person, Whitney Cade, Patrick Hays, Claire Williams, Blair Lehman, and Arthur Graesser. Guru: A computer tutor that models expert human tutors. In *International Conference on Intelligent Tutoring Systems*, pages 256–261. Springer, 2012.

[17] Reinhard Pekrun, Thomas Goetz, Lia M Daniels, Robert H Stupnisky, and Raymond P Perry. Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology*, 102(3):531, 2010.

[18] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98, 2014.

[19] Michael Wixon and Ivon Arroyo. When the question is part of the answer: Examining the impact of emotion self-reports on student emotion. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 471–477. Springer, 2014.

[20] Konstantin Zakharov. Affect recognition and support in intelligent tutoring systems. Master's thesis, University of Canterbury, Computer Science and Software Engineering, 2007.

[21] Ramón Zatarain-Cabada, María Lucía Barrón-Estrada, José Luis Olivares Camacho, and Carlos A Reyes-García. Affective tutoring system for Android mobiles. In *International Conference on Intelligent Computing*, pages 1–10. Springer, 2014.