The Opportunities and Limitations of Scaling Up Sensor-Free Affect Detection

Michael Wixon
Worcester Polytechnic Institute
100 Institute Road
Worcester, Massachusetts
mwixon@wpi.edu

Ivon Arroyo
Worcester Polytechnic Institute
100 Institute Road
Worcester, Massachusetts
iarroyo@wpi.edu

Kasia Muldner
Arizona State University
699 Mill Street
Tempe, Arizona
Katarzyna.Muldner@asu.edu

Winslow Burleson
Arizona State University
699 Mill Street
Tempe, Arizona
winslow.burleson@asu.edu

Cecil Lozano
Arizona State University
699 Mill Street
Tempe, Arizona
calozano@asu.edu

Beverly Woolf
University of Massachusetts-Amherst
140 Governors' Drive
Amherst, Massachusetts
bev@cs.umass.edu

ABSTRACT

We develop and analyze affect detectors for four affective states: confidence, excitement, frustration and interest. We utilize easy to implement self-report based "ground truth" measurements of affect within a tutor, and model them as continuous variables that are later discretized into positive, neutral, and negative valence classifications; this distinguishes our work from detectors which model affective states as binary. We explore the opportunities and limitations of cross validation with regard to potentially distinct sample groups.

Keywords

Affective computing, human factors, intelligent tutors, prediction, models, feature engineering, sensor-free affect detection

1. INTRODUCTION

One key factor that influences students' academic success is their emotions and general affective experience while learning. For instance, positive affect has a facilitative effect on cognitive functioning in general [1], and improved performance on creative problem solving in particular [2, 3]. Moreover, students who are interested in an activity persevere in the face of failure, invest time when needed, and engage in mindful processing [4]. Even some emotions traditionally viewed as negative can be beneficial – for example, confusion is associated with learning under certain conditions [5]. In contrast, the affective state of boredom reduces task performance [6], increases ineffective behaviors such as gaming the system [7], and tends to be persistent once experienced [7].

Given the pivotal role that affect plays in education, both in short-term performance outcomes and in long-term career choices, there is growing interest in developing educational technologies that can recognize and respond to student affect. Here, we focus on the first thrust, namely affect recognition.

The process of modeling motivation and emotion is summarized in Figure 1, which shows how emotions are highly dependent on context, and are expressed in behaviors. Thus, when designing models to assess student emotion, it is essential to empirically understand which factors impact a student's emotional state, and how the affective state is revealed by the student in terms of subsequent actions and behaviors.

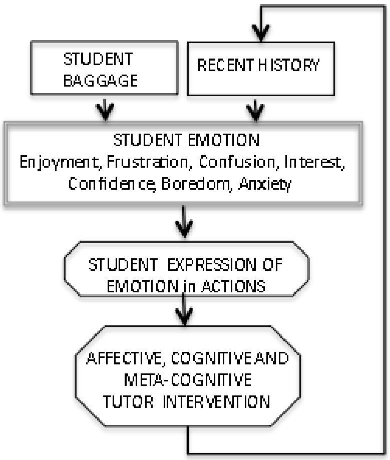

STUDENT BAGGAGE
RECENT HISTORY
STUDENT EMOTION
Enjoyment, Frustration, Confusion, Interest, Confidence, Boredom, Anxiety
STUDENT EXPRESSION OF EMOTION in ACTIONS
AFFECTIVE, COGNITIVE AND META-COGNITIVE TUTOR INTERVENTION

Figure 1: Model of Student Emotion while Learning (arrows indicate dependence, causality, precedence). A student's emotion while learning (grey frame) is originally unknown and hidden. It is influenced by the "student's baggage" (initial achievement, affective predisposition) and recent history of the student in the software (tutor moves or student actions). This article focuses on the top boxes: how student baggage and recent history help to predict current emotional states.

One approach to modeling affect, summarized in a recent review [8], pertains to using sensing devices. For instance, in our past work, we have created models of affect using data from a camera, pressure mouse, skin conductance bracelet, and pressure chair cushions, in conjunction with data coming from a student's interaction with an intelligent tutoring system [9-11]. The subsequent models achieved 85% accuracy when compared to the students' self-reported emotion. Muldner et al. [12] used data from a subset of these sensing devices plus an eye tracker to detect moments of delight during instructional activities. D'Mello et al. [13] used dialog and posture features to model affective states. In Conati's model [14], affect is modeled using one sensor modality, namely an EEG, in addition to interaction features [15].

While this research highlights the utility of sensors for affect recognition, they can not be widely disseminated in schools where the tutoring systems are used, though this may not be true in the future. Data collection is thus more challenging beyond lab studies. Thus, researchers have begun exploring sensor free affective detection. For instance, Baker et al. [16] used only data from students' interaction with a tutor to model affective states such as frustration.

The work reported in this paper adds to research on sensor-free affective models. Specifically, our goal is to better understand contextual predictors of student emotion, and to generate models that use the context in which student emotion occurs to predict this emotion, based on student behaviors within the software. To replace the rich physiological information that sensors provided, we focus on feature engineering, such as summaries of "recent history" of student actions. Additionally, our second goal was understand the utility of students' affective predispositions – attitudes, general values, preferences, and self-efficacy for the domain – for affect detection (see Figure 1). Last but not least, we analyze the generalizability of our affect detectors to different populations of students to other students in new schools.

2. METHODS

2.1 Participants

We used three data sets to train and test ten separate models.

2009 Data Set. An affect detector was built and tested using 295 students, 7th, 8th, 9th and 10th graders from two rural area schools in Massachusetts in the Spring of 2009, using six fold student level batch cross validation [17]. On average, 1138 instances (problem-student interactions) were split across six batches used to train and test each affect model.

2011 Data Set. An affect detector was built and tested using 123 students, 7th and 8th graders from a third rural area school in Massachusetts in 2011, using three fold student level batch cross validation [17]. On average, 120 instances (problem-student interactions) were split across three batches and used to train and test each affect model.

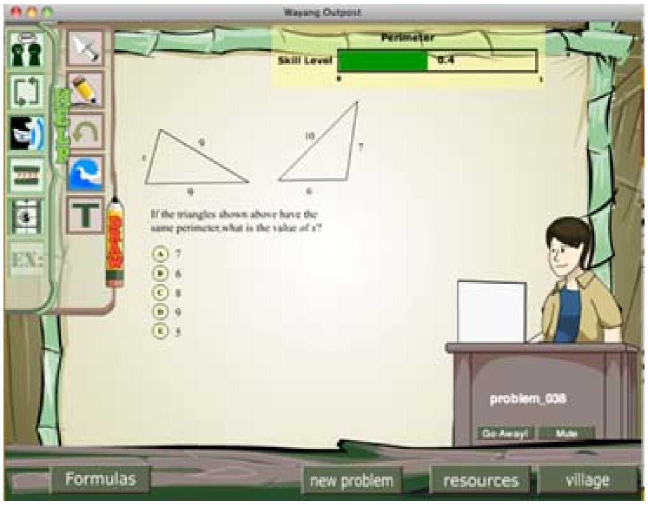
2013 Data Set. An affect detector was built and tested using 43 students, 7th, and 8th graders from two schools in California and Arizona in the Summer of 2013, using two fold student level batch cross validation [17]. On average, 76 instances (problem-student interactions) were split across two batches and used to train and test each affect model.

2.2 Wayang Outpost

The test-bed for this research was Wayang Outpost (see Figure 2). Developed at UMass-Amherst, this tutor shows evidence of promoting effective math learning, has been used by tens of thousands of students in the United States and has consistently shown significant learning gains, e.g., on mathematics tests (an increase of 12% from pre- to post-test after only 4 class periods), and on state standard exams (92%) as compared to students not using Wayang (76%) [11, 18, 19]. Students using Wayang have also improved more on MAP scores compared to control groups (MAP is a national test of Northwest Evaluation Association on specific topics).

2.2.1 Pedagogical Approach

The pedagogical approach of the Wayang Tutor is based on cognitive apprenticeship [20] and mastery learning. Cognitive apprenticeships are designed to bring tacit processes into the open, so that students can observe, enact, and practice them with help from the teacher. This process involves several phases: modeling (introduction to the topic via worked-out examples, making steps explicit, and working through a problem aloud); practice with coaching (offering feedback and hints to sculpt performance to that of an expert's); scaffolding (putting into place strategies and methods to support student learning, offering hints as well as worked-out examples and tutorial videos); and reflection (self-referenced progress charts that allow students to look back and analyze their performance).



If the triangles shown above have the same perimeter, what is the value of x?

A 7
B 6
C 8
D 9
E 5

Figure 2: Learning companions use gestures to offer advice and encouragement. Students can ask for hints or click the "solve it" button. Animations, videos and worked-out examples add to the spoken hints about the steps in a problem.

An important part of cognitive apprenticeship is the provision of materials just beyond what the learners can accomplish by themselves. Vygotsky referred to this as the Zone of Proximal Development (ZPD) and believed that fostering development within this zone leads to the most rapid learning [21]**.** We have operationalized and parameterized ZPD within the context of intelligent tutoring systems [19] and formalized a mechanism for adaptive problem selection that tailors the difficulty of subsequent math problems to past student performance and effort [19]. Wayang also identifies the most critical cognitive skills and predicts the likelihood of success on future problems related to these skills [9]. Wayang supports students by offering hints, examples, short video tutorials, and animations [22-24]. Rich multimedia help is provided when students make mistakes or ask for help, following principles of multimedia learning theory [25].

Teachers can access real-time assessments about individual student progress via the “Teacher Tools”, which allow them to spot and focus on students who need help, problems that are hard for everybody, and math skills with which the class as a whole is struggling.

2.2.2 Affective Learning Companions.

In our past work, we integrated into Wayang gendered and ethnic learning companions (male and female, White, Hispanic and African American), whom offered advice and encouragement by talking to students (see Figure 2 for a sample character). These companions can gesture and train attributions for “success/failure”, e.g., that intelligence is malleable, perseverance and practice are needed to learn, making mistakes is an essential part of learning, and failure is not due to a lack of innate ability. In controlled randomized studies with hundreds of students, certain groups of students (females and students with disabilities) reported decreased frustration and increased confidence levels when working with learning companions and increased frustration when companions were not present [26]. In addition, student enjoyment and interest were higher compared to students not given learning companions, suggesting that such affective pedagogical agents can impact students’ emotions [27, 28]. Moreover, students receiving companions described higher self-efficacy in mathematics, and exhibited more productive behaviors within the tutor.

3. PROCEDURE

In the present study, while working within Wayang Outpost, students were periodically prompted to report their current affective state, using a simple dialogue box. The design of these prompts was based on prior work used to gather information on “the range of various emotional states during learning” [29], where affective states are placed on spectra ranging in valence from negative to positive. The following affective states were measured with a Likert scale (1-5): confidence, excitement, frustration and interest. Each of these scales is bipolar (e.g. confidence/anxiety). For simplicity we will refer to each of these bipolar scales as confidence, excitement, frustration and interest. In this article, a higher Likert score indicates a positive level of the affect in question (i.e., for confidence, 5 is Highly Confident, while 1 is Anxious). In the 2009 and 2011 data sets all four affects were examined, however for the 2013 data set only excitement and interest were measured via self-report.

Recognizing emotion from log data involved a seven step process. First, mathematics problems that students were not expected to solve were removed (e.g., topic introductions and example problems). Second, the student data was batched to ensure each batch had a representative sample of all “ground truth” Likert scale self-reports for all four emotions. Third, missing values were imputed at the batch level using a multiple regression algorithm in SPSS [30], thus filling all cells of missing data with estimate values. Fourth, outliers were identified at the full data set level also using SPSS. Fifth, engineered features were computed from the initial raw log data; some rows of data (e.g. topic introduction problems & example problems where students were meant to observe rather than interact with the system) were removed at this level as well. Sixth, the data was split into ten data sets: one for each combination of year and the four affects to be detected (e.g. confidence 2009, excitement 2013, etc). Seventh, forward feature selection and a linear regression algorithm was run in Rapidminer [31] under batch cross validation [17] in order to build the ten regression models, one for detecting each of the four affects in each of the two sample groups, 2009 and 2011, and two for detecting excitement and interest in the 2013 data set (as only two emotions were self-reported in 2013). Step two (Data Cleaning & Batching), step five (Feature Engineering), and step seven (Model Creation--running the linear regression algorithm) will be addressed in greater detail.

3.1 Data Cleaning & Batching

Data was batched at the student level, meaning that the data from one student could span across more than one batch. The process of batching was not completely random as consideration was given to preserving roughly equal representations of the target self-reported affect in each batch. Thus, students were assigned to batches randomly several times, and each batch was examined to show how many times students had responded with each value of the Likert scale for a given affect. For example, if one batch included 80 instances of students responding with 1 (one) in terms of frustration (low frustration) and another batch included only 10 instances of students with responses of 1 for frustration then that set of batches was rejected and batching was performed again. In some cases it was necessary to manually swap individual students between batches in order to maintain a balanced ratio of responses. The size and quantity of the batches were also limited by concerns of over representation. For example, in the 2011 data set there were only 10 reported cases of interest > 3 out of a total of 105 cases. The fact that less than 10% of our data reported a positive valence in interest for this data set partially explains the relatively poor results of the detector trained on 2011 data, and attempts to “balance” batches by making proportions of each Likert response across batches as equal as possible. It also addresses why the large 2009 data set is split into six batches while the much smaller 2011 data set could only be split into three batches.

3.2 Feature Engineering

The majority of the features were derived from eight low level descriptions of students’ behavior with each problem a student saw (Table 1). Each state first acts as an if-statement predicated that the statement preceding it is not true (e.g. if a student did not SKIP, then the problem is evaluated to see if they met the criteria of NOTR and so on).

These eight simple student states were mutually exclusive and assigned per problem, i.e., for a given problem a student’s actions might be classified as ATT vs. SOF. From these seven features, 21 new features were generated by looking at the prior 3 actions (i.e., NOTRLast3), each of which weighs a more recent instance more heavily than the one that preceded it; for instance, in NOTRLast3 the immediate preceding action is worth 3, the action before that is worth 2, and so on. The remaining features were patters of behaviors, derived from transitions from one student-problem interaction state to another (e.g. NOTR→ATT means that the current student-problem interaction has a state of ATT, and the previous one had a state of NOTR). Due to the fact that several features were based on the prior three actions or prior three transitions between problems, the first four problems of any student’s work within Wayang were excluded from our analyses. This also means that, going forward, these detectors will only be usable after the student has already completed four problems.

Features also included running tallies of incorrect attempts, hints seen, problems solved on first attempt, and other assorted student actions aggregated over the current problem and prior three problems. Several hundred features were generated and only a small number were selected for use in models in this work; we limit our discussion to the features that were selected.

Table 1. Eight Low Level Student States

Student State	**Description of student Behavior**
SKIP	The student did nothing and skipped the problem.
NOTR (Not Reading)	The student made a first attempt to solve a problem in a time under 4 seconds –not enough time to even read the problem.
GIVEUP	The student took some action, but then skipped the problem without solving it.
SOF (Solved on First Attempt)	The student solved the problem on their first attempt, without seeing any help.
BOTT (Bottom Out Hint)	The student saw all hints available, including the last available hint that gave the answer.
SHINT (Student Hint Request)	Student answered the math problem eventually right, with at least 1 hint.
ATT (Attempt)	The student didn't see any hints and solved it correctly after 1 wrong attempt.
GUESS	The student solved it correctly with no hints and more than 1 incorrect attempt.

For the features shown in Table 2, "Avg" denotes an average taken across the prior four problems, "Last4" denotes the sum of the prior four problems, "Max" denotes the maximum number of actions in a given problem over the prior four problems, "Min" denotes the minimum number of actions in a given problem over the prior four problems, and % denotes the ratio of a particular action in the past four problems over the total actions in the past four problems.

3.3 Model Creation

Once the batching of the data was finalized, each data set was split into the four subsets, each addressing the emotion in question: confidence, excitement, frustration, and interest. Initially, forward feature selection (with a limit of ten features) was carried out for each of the four types of affect for each data set, with student-level batch cross validation [17].

Linear regression was performed in Rapidminer [31] on each of these new subsets under batch cross validation [17]. The models were assessed by Pearson's R to determine their correlation with the target affect. Further, in order to create a discrete classification measure of affect, the Likert scale responses and linear regression model output were rounded to the nearest integer and then discretized as follows: All responses below 3 on the Likert scale were labeled as "negative", all responses equal to 3 were labelled "neutral", and all responses above 3 were labeled as positive. These classification results were assessed using weighted kappa [32], which is a measure of agreement for polynomial classified targets. Similarly to typical Cohen's kappa [33], a zero denotes agreement due to random chance, while a one denotes perfect agreement between the model and student self-reports of affect.

While detector results obtained under batch cross validation should guard against overfitting, there is still the potential risk that the results may be overfit to the sample group used in the study. In particular, even with batch cross validation, all the batches are drawn from the same sample group, who may share various specific traits. Therefore, the batch cross validated models trained using the 2009 data set was applied to the 2011 & 2013 data sets and vice-versa. This was done to provide a more conservative estimate of the models' generalizablity to new data sets, given that the samples were collected from distinct groups of students at distinct points in time.

Table 2. Features from Students' Interaction in Wayang

AvgTimeToSolve – The average of time to solve a problem.
LogTimePerAction – The logarithm $\log_{10}$ of the time per action
AvgTimePerAction – The average time per action
Hints – Total hints given on current problem
Wrong – Total wrong attempts on the prior problem
WrongLast4 – Total wrong attempts aggregated over the current and last 3 problems.
MaxWrong – The maximum number of incorrect attempts
MaxActions – The maximum number of actions
MinWrong – The minimum number of incorrect attempts
TimetoSolve – Time to solve a problem
LogTimeInTutor – Logarithm $\log_{10}$ of student's time in tutor.
TimeInTutor – Total student's time in tutor.
MinTimePerAction – The minimum time per action of the past 4 problems.
MinLogTimePerAction – The minimum of the logarithm $\log_{10}$ of seconds per action.
TotalActions – The total actions of the prior problem.
%Wrong – The percent of incorrect attempts.

4. RESULTS

4.1 Feature Selection

Forward feature selection yielded a total of forty eight features. These features were split across ten different detectors/models, four for the 2009 data set, four for the 2011 data set, and two for the 2013 data set where only self-reports on excitement and interest were collected. While there were ten models and ten features used per model, only 48 features were required rather than 100, because some features were used in more than one model. Twenty seven of these features were engineered from the states described in Table 1. Of the remaining twenty one features, sixteen were based on other student interactions within the system (see Table 2). Many of these features were based on student actions on an immediate given problem, but some denoted with "Avg", "Max", "Min" or "%" are based upon the current problem and three preceding problems: "Avg" denoting Average, "Max" denoting maximum, "Min" denoting minimum, and "%" denoting the percentage of a particular action out of the total actions taken over the current and prior three problems.

The remaining five features (see Table 3) were based on students' responses on the pretest, surveys, and the experimental conditions. These features remain constant from problem to problem.

Table 3. Pretest and Agent Based Features

Features Based on Survey Responses and Agent's Behavior
CON – Baseline measure of confidence when problem solving.
FRUS – Baseline measure of frustration when problem solving.
INT – Baseline measure of interest towards problem solving.
MathValuing – Baseline measure of the degree to which the student values mathematics.
pre_lor –Student's mastery orientation (willingness to learn new and interesting things in spite of challenge) based on a survey.

4.2 Model Performance

The R values of the linear regression models derived from the selected features achieved a fit comparable with prior work detecting frustration [34, 35], as well as boredom, confusion, and flow [35]. Specifically, prior work has achieved detectors of frustration with kappa values ranging from 0.16 to 0.32 [16], and boredom at kappa = 0.28 [16]. While the detectors presented in this paper may achieve slightly lower kappas than detectors presented in the above cited work, it's important to note that our kappas are weighted [32], which suffer a penalty as compared to the typical Cohen's kappa [33] that is meant for bivariate classification. Consequently our model distinguishes between three possible classifications rather than two. This increased the likelihood of accidental misclassification, but with the benefit of more sensitive measurement. One cost of modeling affect as polynomial rather than binary is that binary classification has metrics for false and true positive and negative rates such as sensitivity and specificity [36] or A' [37], which we cannot utilize in this work.

It is important to note the sample size when considering the relative strength of each model. As previously mentioned the largest sample was found in the 2009 data set, where each model was built on an average of 1138 instances split across six batches. The 2011 data set contains 120 instances split across three batches. However, for the 2011 data set there were only ten instances of positively valenced interest. The particularly low values of interest in 2011 may explain why the 2009 derived model better predicts interest in that sample than the 2011 derived model.

Tables 4 through 7 show performance indicators of each model, which consist of R values (indicating model fit) and weighted kappas [32] (denoted by "K", indicating classification power into low/neutral/high levels). Each cell contains performance results for a model created from a dataset indicated by the column, and evaluated over a dataset indicated by the row. Note that values along the diagonal (in bold) correspond to testing and training over the same data set. In such cases, student level batch cross validation was used to prevent overfitting. The process of applying the model to the same data set (to generate estimates of the emotion) is thus slightly different than for other cells. Under batch cross validation, a separate model is generated (i.e. trained) for each batch, and estimations/classifications are made for the testing batch. The performance of six distinct models is thus aggregated in the end for the 2009 data set; the performance of three distinct models is aggregated in the 2011 data set); and the performance of two distinct models is aggregated in the case of the 2013 data set.

Table 4. Confidence Detector Performance (Pearson's R & Cohen's Kappa)

	2009 Model	2011 Model

2009 Data Set
N = 1102
	R = 0.404
	K = 0.200
		R = 0.306
		K = 0.163

2011 Data Set
N = 127
	R = 0.515
	K = 0.249
		R = 0.238
		K = 0.147

Table 5. Frustration Detector Performance (Pearson's R & Cohen's Kappa)

	2009 Model	2011 Model

2009 Data Set
N = 1159
	R = 0.372
	K = 0.173
		R = 0.307
		K = 0.146

2011 Data Set
N = 125
	R = 0.374
	K = 0.139
		R = 0.341
		K = 0.281

Table 6. Excitement Detector Performance (Pearson's R & Weighted Kappa)

	2009 Model	2011 Model	2013 Model

2009 Data
N = 1145
	R = 0.224
	K = 0.151
		R = 0.211
		K = 0.083
			R = -0.089
			K = -0.022

2011 Data
N = 122
	R = 0.454
	K = 0.278
		R = 0.316
		K = 0.131
			R = -0.142
			K = -0.050

2013 Data
N = 66
	R = 0.004
	K = 0.102
		R = 0.201
		K = -0.024
			R = 0.137
			K = 0.192

Table 7. Interest Detector Performance (Pearson's R & Weighted Kappa)

	2009 Model	2011 Model	2013 Model

2009 Data
N = 1145
	R = 0.240
	K = 0.090
		R = 0.058
		K = 0.026
			R = 0.071
			K = -0.024

2011 Data
N = 105
	R = 0.300
	K = 0.140
		R = 0.174
		K = 0.005
			R = -0.001
			K = -0.036

2013 Data
N = 86
	R = 0.006
	K = 0.055
		R = 0.153
		K = -0.023
			R = 0.020
			K = -0.144

In general, the results in Tables 4-7 show that: a) affect detectors for confidence/anxiety, excitement and frustration achieve reasonable levels of performance, while for interest/boredom, the R and Kappa values are much lower; b) models generated over larger datasets transfer better to smaller datasets, compared to the other way round; c) models perform similarly well across 2009 and 2011 but not as well over the 2013 dataset, which corresponded to a summer camp in a different part of the country; d) models created over the 2013 dataset don't transfer well to the 2009-2011 datasets either. These points will be explored in the discussion section.

4.3 Linear Regression Models

The linear regression models for the four affect states are displayed in Tables 8 through 11.

Table 8. Models of Confidence

2009 Features	Weight	2011 Features	Weight
NOTR→BOTT	-53.00	GIVEUPLast3	75.77
BOTT→GUESS	-21.64	NOTR→BOTT	-40.42
GIVEUPLast3	-10.74	BOTT→BOTT	5.14
SOFLast3	0.34	SOF→BOTT	-4.96
Pre_LOR	0.34	SOFLast3	1.06
MinLogTimePerAction	0.31	Pre_LOR	0.87
Wrong	-0.20	MaxWrong	0.28
WrongLast4	-0.07	WrongLast4	-0.27
FRUS	-0.04	CON	0.10
CON	0.04	TimetoSolve	0.01

Table 9. Models of Frustration

2009 Features	Weight	2011 Features	Weight
NOTR→NOTR	-99.37	GUESS→NOTR	-79.74
GIVEUP	11.56	SHINT→NOTR	-36.07
GUESS→SOF	-2.47	GIVEUP	-22.85
SHINT→SOF	-1.58	SHINT	-3.32
%Wrong	0.66	SOF	-1.77
AvgTimePerAction	-0.24	%Wrong	0.98
WrongLast4	0.09	Pre_LOR	-0.53
TotalActions	0.05	INT	-0.12
FRUS	0.04	CON	-0.09
INT	-0.04	MaxActions	0.08

Table 10. Models of Excitement

2009 Features	Weight	2011 Features	Weight	2013 Features	Weight
BOTT→ NOTR	-74.00	GIVEUP	35.24	SHINT→ BOTT	66.89
SOF→ NOTR	-22.52	BOTT→ SHINT	25.32	SHINT→ SKIP	-4.31
Min Wrong	-2.57	Pre_LOR	-0.84	SKIP	2.80
SOF→ BOTT	2.55	Hints Seen	-0.49	SHINT→ SHINT	2.09
Incorrect Attempts	0.14	INT	-0.14	Pre_LOR	-0.76
INT	-0.14	Wrong Last4	0.05	Hints Seen	-0.36
Wrong Last4	0.12	CON	0.05	CON	0.08
Max Wrong	-0.07	LogTime InTutor	-0.04	AvgTime ToSolve	0.01
MinTime PerActio	-0.01	AvgTime PerAction	-0.01	TimeIn Tutor	< 0.01
TimeIn Tutor	< 0.01	AvgTime ToSolve	< 0.01		

Table 11. Models of Interest

2009 Features	Weight	2011 Features	Weight	2013 Features	Weight
SOF→ SHINT	1.16	GIVEUP	349.31	NOTR→ SOF	30.20
SHINT	1.06	GIVEUP →SOF	-180.20	BOTT	19.68
%Wrong	-0.56	SHINT→ NOTR	52.42	SOF→ GUESS	-8.15
SOF	0.41	SHINT→ SHINT	26.43	SKIP→ SOF	-7.33
Pre_LOR	0.37	NOTR→ SOF	-17.61	SHINT →GUES	6.43
INT	0.08	SOF→ NOTR	17.00	LogTime InTutor	-0.09
Total Actions	-0.05	BOTT→ BOTT	7.14	Max Wrong	-0.07
MinTime PerActio	-0.02	Math Valuing	0.09	INT	0.05
TimeIn Tutor	< 0.01	MinTime PerAction	0.03	TimeIn Tutor	< 0.01
		LogTime InTutor	0.02		

5. DISCUSSION

In this paper, we have proposed several models of affect based on students' interaction with a tutoring system. In so doing, we have independently replicated prior work on sensor-free affect detection and contributed to existing work on predictive features of student affect and methods for building models of affect. In the following section we address opportunities and challenges regarding generalizability of the models to new populations.

A major opportunity is to develop detectors which respond to differences between classrooms, schools, and even different regions of the country. We generated a rich set of features which combined student behaviors in the last problem seen, recent history, patterns of student behaviors, and even students' affective background before starting the tutoring session. A combination of features from all these categories were best predictors for each affective state, showing that a variety of student descriptors as well as their behaviors can help to predict emotional states while learning.

It is important to note that while some of the features we used bear a similarity to those in other research, the features are dependent on the environment from which they are inferred. Thus, validation is needed to ensure that these features transfer and apply to other tutoring systems, such as Wayang Outpost.

In designing the features used, consideration was given to other detectors of affect [16, 38]. There is a tension between trying to use similar features from other systems, and recognizing features as being contextually distinct; this makes detector construction a custom work on each system. In the future, it is our hope to design even more informative features. This could be done by examining the data to look for patterns of behavior that align to affective states, and to observe students using the software for behaviors that might have been overlooked and could be indicators of affect. While examining the data in such a way could "pollute" a researcher's perspective and result in features that may overfit to a particular data set, this may be a necessary build generalizable detectors.

Much of our feature selection work relied on the atheoretical approach of simple forward selection that yielded some features that may be only coincidentally correlated with our target affects. The best way to increase fidelity in identifying which features are true expressions of an affective state is to examine which coefficients remain similar in sign and magnitude across detectors built for different data sets. For example, in both confidence models generated, NOTR→BOTT enters into the regression model with a negative coefficient. This means that transitioning from responding to a problem in under four seconds to using a bottom out hint is negatively correlated with confidence, in both models generated over different data sets. Both of these behaviors seem expressions of disengagement, and other potentially disengaged student states like GIVEUP and GUESS also figure largely into both models. Unfortunately, the similarity in these states (as expressions of disengagement) may make the models more different than they need to be as in the case of NOTR→NOTR versus GUESS→NOTR in the case of frustration.

The statistical power of using a larger and therefore likely more diverse data set is evident from our findings. In all cases (with the exception of frustration), the 2009 model outperforms the 2011 when applied to the 2011 data set. The fact that the 2009 data set has about twice as many participants and roughly ten times as many affect reports may explain this trend. Thus, a larger and more diverse data set seems to generalize better to new samples and groups of students.

Finally, it's worth noting that the 2013 models transferred poorly to 2009 and 2011 datasets, and that the 2013 data set came from summer school students from the southwestern United States (Arizona & California). Models trained on the 2009 or 2011 data sets do not appear to generalize to the 2013 data set, or vice versa. We believe this is because the 2013 dataset was unique in several ways: it came from a different region of the country; it corresponded to students working in a summer program as opposed to during a typical school year; a slightly different version of Wayang Outpost was used. In addition, the 2013 students only self-reported on two affective states: excitement and interest, but not confidence or frustration. While batch cross validation may address within sample distinctness between participants, it does little to address how well the model will perform when applied to a distinct new sample group whose participants are distinct from the training group (e.g. summer school vs. not summer school, within a regular math class).

Limitations of generalizability across samples might be the largest challenge, also found in other work. In a recent study [39], detectors trained on student sample groups from urban, suburban, and rural areas were shown to have difficulty generalizing to a different sample group. For example, a detector of Confusion trained on suburban students under batch cross validation achieved a kappa of 0.38 when applied to suburban students, but performed at chance when applied to rural students with a kappa of 0, and only slightly better when applied to urban students with a kappa of 0.06 [39]. This shows that while cross validation may provide a conservative estimate on how well a model may generalize to new data, the accuracy of this estimate is conditioned upon the training data being representative of the population to which the model is to be applied to.

6. ACKNOWLEDGEMENTS

This research was funded by the National Science Foundation, #1324385, Cyberlearning DIP, Impact of Adaptive Interventions on Student Affect, Performance, and Learning; Burleson, Arroyo and Woolf (PIs). Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

7. REFERENCES

[1] Hidi, S. (1990) Interest and Its Contribution as a Mental Resource for Learning. *Review of Educational Research*. 60(4).

[2] Isen, A.M., K. Daubman, and G. Nowicki (1987) Positive affect facilitates creative problem solving. *Journal of Personality and Social Psychology*. 52, 1122–1131.

[3] Pekrun, R., A.J. Elliot, and M.A. Maier (2009) Achievement Goals and Achievement Emotions: Testing a Model of Their Joint Relations With Academic Performance. *Journal of Educational Psychology*. 101(1), 115–135.

[4] Lepper, M. (1988) Motivational Considerations in the study of Instruction. *Cognition and Instruction*. 5(4), 289-309.

[5] D'Mello, S.K., B. Lehman, R. Pekrun, and A.C. Graesser (in press) Confusion Can be Beneficial For Learning. *Learning & Instruction*.

[6] Pekrun, R., T. Goetz, L. Daniels, R. Stupinsky, and R. Perry (2010) Boredom in Achievement Settings: Exploring Control–Value Antecedents and Performance Outcomes of a Neglected Emotion. *Journal of Educational Psychology*. 102(3), 531-549.

[7] Baker, R.S.J.d., S.K. D'Mello, M.M.T. Rodrigo, and A.C. Graesser (2010) 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.

[8] Calvo, R.A. and S. D'Mello (2010) Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE transactions on affective computing*. 1(1), 18-37.

[9] Cooper, D.G., K. Muldner, I. Arroyo, B.P. Woolf, W. Burleson, and R. Dolan (2010) Ranking Feature Sets for Emotion Models used in Classroom Based Intelligent Tutoring Systems. *In Proceedings of International Conference on User Modeling and Adaptive Presentation (UMAP'10)*, 135-146.

[10] Cooper, D., I. Arroyo, B. Woolf, K. Muldner, W. Burleson, and R. Christopherson (2009) Sensors Model Student Self Concept in the Classroom. *In Proceedings of UMAP 2009, First and Seventeenth International Conference on User Modeling, Adaptation and Personalization*, 30-41.

[11] Arroyo, I., D.G. Cooper, W. Burleson, B.P. Woolf, K. Muldner, and R. Christopherson (2009) Emotion Sensors Go To School. *In Proceedings of 14th International Conference on Artificial Intelligence in Education (AIED'09)*, 17-24.

[12] Muldner, K., W. Burleson, and K. VanLehn (2010) "Yes!": Using Tutor and Sensor Data to Predict Moments of Delight

during Instructional Activities. *In Proceedings of User Modeling, Adaptation, and Personalization*, 159-170.

[13] D'Mello, S. and A. Graesser (2007) Mind and Body: Dialogue and Posture for Affect Detection in Learning Environments. *In Proceedings of International Conference on Artificial Intelligence in Education*, 161-168.

[14] Conati, C. and X. Zhou (2002) Modeling Students' Emotions from Cognitive Appraisal in Educational Games. *In Proceedings of ITS 2002, 6th International Conference on Intelligent Tutoring Systems.*,

[15] Conati, C. and H. Maclaren (2009) Modeling User Affect from Causes and Effects. *In Proceedings of UMAP 2009, First and Seventeenth International Conference on User Modeling, Adaptation and Personalization*, 10 pages.

[16] Baker, R.S.J.d., S.M. Gowda, M. Wixon, J. Kalka, A.Z. Wagner, A. Salvi, V. Aleven, G. Kusbit, J. Ocumpaugh, and L. Rossi (2012) Sensor-free automated detection of affect in a Cognitive Tutor for Algebra. *In Proceedings of 5th International Conference on Educational Data Mining*, 126-133.

[17] Efron, B. and G. Gong (1983) A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician*. 37, 36-48.

[18] Arroyo, I., H. Mehranian, and B. Woolf (2010) Effort-based tutoring: An empirical approach to intelligent tutoring. *In Proceedings of 3rd International Conference on Educational Data Mining*, 1-10.

[19] Beal, C.R., R. Walles, I. Arroyo, and B.P. Woolf (2007) On-line tutoring for math achievement testing: A controlled evaluation. *Journal of Interactive Online Learning*. 6(1), 43-55.

[20] Collins, A., J.S. Brown, and S.E. Newman (1989) *Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics*, in *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser* L.B. Resnick, Editor Lawrence Erlbaum Associates: Hillsdale, NJ. p. 453-494.

[21] Vygotsky, L. (1978) *Mind in society*: Harvard University Press.

[22] Arroyo, I. and B. Woolf (2005) Inferring learning and attitudes from a Bayesian Network of log file data. *In Proceedings of 12th International Conference on Artificial Intelligence in Education (AIED'05)*, 33-40.

[23] Arroyo, I., C. Beal, T. Murray, R. Walles, and B.P. Woolf (2004) Web-Based Intelligent Multimedia Tutoring for High Stakes Achievement Tests Intelligent Tutoring Systems. *In Proceedings of 7th Internatinal Conference on Intelligent Tutoring Systems (ITS'04)*, 142-169.

[24] Arroyo, I., K. Ferguson, J. Johns, T. Dragon, H. Meheranian, D. Fisher, A. Barto, S. Mahadevan, and W. B.P. (2007) Repairing disengagement with non-invasive intervention. *In Proceedings of 13th International Conference on Artificial Intelligence in Education*, 195-202.

[25] Mayer, R.E. (2001) *Multimedia Learning* New York: Cambridge University Press.

[26] Arroyo, I., W. Burleson, M. Tai, K. Muldner, and B. Woolf (in press) Gender Differences In the Use and Benefit of Advanced Learning Technologies for Mathematics. *Journal of Educational Psychology*.

[27] Woolf, B., I. Arroyo, K. Muldner, W. Burleson, D. Cooper, R. Dolan, and R. Christopherson (2010) The Effect of Motivational Learning Companions on Low Achieving Students and Students with Disabilities. *In Proceedings of The 10th International Conference on Intelligent Tutoring Systems (ITS'10)*, 327-337.

[28] Arroyo, I., B.P. Woolf, J.M. Royer, M. Tai, K. Muldner, W. Burleson, and D. Cooper, *Gender Matters: The Impact of Animated Agents on Students' Affect, Behavior and Learning*, in *Technical report UM-CS-2010-020*2010, UMASS Amherst.

[29] B. Kort, R.R. and R.W. Picard (2001) An Affective Model of Interplay Between Emotions and Learning: Reengineering Educational Pedagogy-Building a Learning Companion. *In Proceedings of International Conference on Advanced Learning Technologies*,

[30] IBM (2012) *IBM SPSS Statistics for Windows, Version 21.0* Armonk, NY: IBM Corp.

[31] Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler, T. (2006) YALE: Rapid Prototyping for Complex Data Mining Tasks. *In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06)*, 935-940.

[32] Cohen , J. (1968) Weighted kappas: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*. 70, 213-220.

[33] Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 20(1), 37–46.

[34] Rodrigo, M. and R. Baker (2009) Coarse-Grained Detection of Student Frustration in an Introductory Programming Course. *In Proceedings of ICER 2009: the International Computing Education Workshop*,

[35] D'Mello, S., S. Craig, A. Witherspoon, B. McDaniel, and A. Graesser (2008) Automatic Detection of Learner's Affect from Conversational Cues. *User Modeling and User-Adapted Interaction*. 18(1-2), 45-80.

[36] Altman, D.G. and M. Bland (1994) Diagnostic tests 1: sensitivity and specificity. *BMJ*. 308, 1552.

[37] Hanley, J. and B. McNeil (1982) The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*. 143, 29-36.

[38] Pardos, Z.A., R.S.J.d. Baker, M.O.C.Z. San Pedro, S.M. Gowda, and S.M. Gowda (2013) Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. *In Proceedings of 3rd International Conference on Learning Analytics and Knowledge*, 117-124.

[39] Ocumpaugh, J., R. Baker, S. Gowda, N. Heffernan, and C. Heffernan (in press) Population validity for Educational Data Mining models: A case study in affect detection. *British Journal of Educational Technology*.