Time to Scale: Generalizable Affect Detection for Tens of Thousands of Students across an Entire School Year

Stephen Hutt
Institute of Cognitive Science
University of Colorado Boulder
Boulder, Colorado, USA
stephen.hutt@colorado.edu

Joseph F. Grafsgaard
Institute of Cognitive Science
University of Colorado Boulder
Boulder, Colorado, USA
joseph.grafsgaard@colorado.edu

Sidney K. D'Mello
Institute of Cognitive Science
University of Colorado Boulder
Boulder, Colorado, USA
sidney.dmello@colorado.edu

ABSTRACT

We developed generalizable affect detectors using 133,966 instances of 18 affective states collected from 69,174 students who interacted with an online math learning platform called Algebra Nation over the entire school year. To enable scalability and generalizability, we used generic interaction features (e.g., viewing a video, taking a quiz), which do not require specialized sensors and are domain- and (to a certain extent) system-independent. We experimented with standard classifiers, recurrent neural networks, and genetically evolved neural networks for affect modeling. Prediction accuracies, quantified with Spearman's rho, were modest and ranged from .08 (for surprise) to .34 (for happiness) with a mean of .25. Our model trained on Algebra students generalized to a different set of Geometry students (n = 28,458) on the same platform. We discuss implications for scaling up affect detection for affect-sensitive online learning environments which aim to improve engagement and learning by detecting and responding to student affect.

CCS CONCEPTS

• Human-centered Computing → **Human Computer Interaction(HCI);** Applied Computing → e-learning

KEYWORDS

Sensor-free, affect detection, machine learning, online education

ACM Reference format:
Hutt, Stephen, Grafsgaard, Joseph F & D'Mello, Sidney K. 2019. Time to Scale, Generalizable Affect Detection for Tens of Thousands of Students across an entire School Year. In *2019 CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glagsow, Scotland, UK.* ACM, New York, NY, USA. 12 pages. https://doi.org/10.1145/3290607.3300726

CHI 2019, May 4-9, 2019, Glasgow, Scotland, UK.

ACM ISBN 978-1-4503-5970-2/19/05...$15.00.
DOI: https://doi.org/10.1145/3290605.3300726

1 INTRODUCTION

Imagine you are tutoring two students in algebra. While working through a problem set for balancing equations you notice that one student has become frustrated with the task, but the other is engaged and eager for more work. You alter your tutoring strategy based on the sensed affect – perhaps giving the frustrated student a hint while ramping up challenge for the engaged student. This level of adaptivity was only possible because you were able to sense your students' affective states – a key dimension of adaptivity given the critical role of affect in learning [20, 48]. In fact, two decades ago, in a popular book called *Emotional Intelligence* [31], Goleman claimed that expert teachers are very adept at recognizing and addressing the affective states of their students. But what these expert teachers see, and how they decide on a course of action, is still an open question despite considerable progress on basic research on affect and learning in classrooms [7, 17, 44].

It is also important to realize that modern learning occurs outside of traditional or even computer-enabled classroom. Massive Open Online Courses (MOOCs) and other online learning environments present a new paradigm for learning. Despite posing new challenges for educators, online learning environments have become an increasingly popular method for e-learning and distance learning [39]. For example, MyMaths [43], a subscription service where teachers can assign tasks to students, review their progress, and provide feedback, is currently used by over 4 million students a year across more than 70 countries. Online learning environments also play a role in traditional environments as alternate instructional paradigms, such as flipped classrooms and other forms of blended learning [55].

Despite their widespread use, it is widely acknowledged that learning with online platforms can be a cold and detached experience [41], with the high disengagement and

dropout rates in MOOCs providing an exemplary case [18, 38]. In contrast, meaningful learning is a deeply emotional experience [9, 48, 49]. Is it possible to augment online learning environments with mechanisms to sense and respond to student affect [7] in order to promote engagement and learning?

Because an intelligent affect-sensitive online learning environment must first sense affect before it can respond to it, researchers in the field of affective computing have spent the last two decades developing systems that can automatically detect affect [7, 17, 25]. However, despite considerable progress, as evident from the literature review (Section 2), current affect detectors are limited in several fundamental ways — e.g., they are trained and validated on data collected from small numbers of students (< 200), across short learning sessions (< 2 hours), usually in a laboratory setting (but see [7, 14, 17, 60] for some exceptions). Further, much of the current work on affect detection does not lend itself to online learning environments, where physical sensors often meet political resistance (e.g., cameras in schools) and financial barriers to scalability (e.g., physiological sensing is still cost prohibitive).

We argue that the time has come for next-generation detectors, which aim at scalability to hundreds of thousands of diverse students across extended time frames of an entire school year and beyond. We ground these ideas in a project involving automated detection of 18 affective states in 69,174 high-school students who use an online math learning platform for an entire school year. To promote generalizability and scalability, we use generic activity features (number of video plays, forum posts, questions complete etc.) that are seamlessly collected as students interact with an the learning platform. We also show that our affect detection models developed in one domain (Algebra 1) can be applied to another (Geometry) without retraining and with minimal loss in accuracy.

2 RELATED WORK

Because the field of affect detection is vast (see reviews in [7, 17, 25]), we focus on affect detection during learning with technology, with a particular emphasis on studies conducted in authentic educational contexts. This work can be divided into sensor-based and interaction-based (sensor-free) affect detection.

2.1 Sensor-based Affect Detection

Much of the work in affect detection has relied on sensing streams beyond the standard input devices. Facial features, acoustic-prosodic cues, posture and body language, physiology and, brain imaging have all been used to detect affective states (see review articles [17, 28]). Sensor-based models have been used to infer a variety of affective states, including frustration [29, 32], confusion [45], engagement [11, 34], interest [24, 42]. Whereas the use of sensors lends itself to laboratory environments [37, 40, 42], classrooms presents a far noisier environment with less experimenter control.

Nevertheless, about a decade ago, researchers began deploying affect sensors in authentic learning contexts. For example, Arroyo et al [5] collected data from 38 high school students and 29 undergraduates as they interacted with an Intelligent Tutoring System (ITS) in a computer-enabled classroom. Using data from physiology sensors, webcams, pressure sensors, and interaction data (i.e., log files from the ITS), they attempted to build detectors of frustration, interest, confidence and excitement. Their research was pioneering despite issues with large sensing data loss and lack of evidence on their generalizability to new students.

Taking a different approach, Hutt et al. [34] harnessed eye gaze to detect mind wandering (related to boredom and engagement [57]) while 135 students interacted with an ITS in their regular biology classroom. Despite the noisy classroom environment, they could obtain valid eye gaze for 95% of the sessions, suggesting that consumer-grade eye tracking could be used in classrooms.

In another classroom study, Bosch et al. [12] used webcam videos (facial expression and body movement) to detect boredom, engaged concentration, confusion, frustration and delight as 137 students played a physics game across two days. They found that facial features could be used to detect affective states, but with considerable data loss (35%) when the face could not be reliably tracked in the video.

Finally, DeFalco et al. [29] use posture and body movement to detect five affective states while 101 students engaged with a military training simulation, designed to improve combat medicine. They utilized data from a depth-sensing camera (Microsoft Kinect) to develop detectors for boredom, confusion, concertation, frustration and surprise.

The variety of sensors used in these studies demonstrates the multiple avenues available for affect detection in computer-enabled classrooms. However, each comes with an additional financial cost and setup cost, limiting how sensor-based detectors might scale more broadly. And although webcams are standard in modern computers, there are several privacy concerns involved in

recording video. These concerns are exacerbated when minors engage with online learning platforms in the privacy of their homes. Interaction-based detectors avoid these issues by providing a sensor-free approach to affect detection, as detailed below.

2.2 Interaction-based Affect Detection

A student interacting with a digital learning environment leaves a rich data trail stored in log files. The logs document student actions, topics covered, videos watched, student preferences, and so on. Interaction-based affect detectors analyze such log data without the need for any sensors beyond standard input devices (mouse, keyboard, etc.). Table 1 lists a representative sample of studies in which interaction-based affect detection was developed and tested for different learning environments. A more in-depth discussion of many of these studies can be found in [7]. Additional studies follow the same trends (e.g., Pardos et al.[45] trained their detectors with data from 229 students)."

These studies all found (to varying degrees of success) that interaction data can be mined to build sensor-free affect detectors. Importantly, most of the studies utilized student-level validation methods (data from the same student can be in the training or test set but not both), which increases the likelihood that the models will generalize to new students.

We also note a number of other commonalities among studies. First, studies have typically considered only a small subset of affective states, at most seven, with an emphasis on concentration, confusion, and frustration. Second, many of the existing studies have relied upon small numbers of students to train their detectors, often due to constraints regarding obtaining a ground truth affect labels. In particular, many studies have relied upon online observations, requiring considerable researcher time to obtain labeled data. Third, much of the existing work has taken place either in the lab or in the classroom. Classrooms present a rich ecologically valid environment however, this presents a limitation as to how many students can be involved in the study as well as how long the study can go on. It is very costly to continue to return to the school for data collection, placing a logistical and financial obstacle to scaling these approaches. As a result most of the studies consist of short learning sessions over one or two days (with a couple of notable exceptions).

3 CONTRIBUTION & NOVELTY OF CURRENT STUDY

The novelty of this contribution is four-fold. First, we collected data from a an online learning platform (Algebra Nation [3]) used across an entire U.S. state, giving us the potential to scale up previous work on affect detection to hundreds of thousands of students. Indeed, whereas previous work has focused on data collected from small numbers of students in one or more classrooms over short periods of time (see Table 1), we collected 133,966 instances of 18 affective states from 69,174 students in 1,898 schools over an entire school year (36 weeks).

Second, whereas previous work on interaction-based affect detection has focused on system-specific activity features (e.g., watching video X while viewing lesson Y or answering question id 4567 correct [7]), resulting in models that are hyper-tuned to a particular learning domain and learning technology. As such, one essentially needs to start over when moving from one platform (or even domain) to another. In contrast, we use generic activity features (e.g., viewing any video lecture, taking any quiz, viewing the discussion board), which are more domain- and platform-independent. Our features are also common to many online learning environments, improving the potential for generalizing our results to other environments. Indeed, we show that not only do our models trained on one domain (Algebra I) generalize to new students within that domain, they also generalize to new students from a different domain (Geometry I).

Third, due to the large scale of data collection, ours is truly an "in the wild" study compared to previous work on affect detection which was often done in the lab or on small homogeneous samples. With Algebra Nation, we have no control over when or how students use the system. Some interactions may be from home, others from the classroom. Some are student-driven others are teacher-led. This large heterogeneous dataset presents a significant challenge for affect detection, but improves the potential to generalize across different use cases – an important criterion of scaling up.

Finally, we consider a large range of 18 affective states, in contrast to previous work focusing on about 7 states. This was needed to accommodate the rich affective profile expected when diverse students engage in learning across extended time frames. It also provides an opportunity to explore how different affective states might be manifested in interaction patterns and which are easier/harder to detect from our generic activity features.

Table 1. Sensor Free Affect detection in Learning Studies

Study	Num. Students	Affective States	Learning Environment	Features	Context	Session Time	Study Duration	Cross Validation	Ground Truth
Ai et al. (2008)[2]	20	Confusion	Why2Atlas	NLP Features	Lab	~ 1 Hour	1 Day	10-fold Instance Level	Retrospective Human Coding
Baker et al. (2012) [6]	89	Concentration, Confusion, Frustration Boredom	Cognitive Tutor Algebra	Interaction Features	Classroom	~ 1 Hour	1 Day	6-fold student-level	Observations
Botelho, Baker, and Heffernan (2017) [14]	646	Concentration, Confusion, Frustration Boredom	ASSISTments	Interaction Features, Student Performance	Classroom Study	~ 1 Hour	1 Day	5-fold Student-level	Observations
Conati and Maclaren (2009) [21]	66	Joy, Distress	Prime Club	Interaction Features, Personality	Classroom	10 Minutes	3 Days	Leave Several Students Out	Self-report Survey
D'Mello et al. (2008) [23]	28	Confusion, Flow, Frustration, Neutral state	AutoTutor	Interaction Features, Context	Lab	32 Minutes	1 Day	10-fold Instance level	Retrospective Human Coding
Lee et al. (2011) [36]	149	Confusion	BlueJ Programming Environment	Interaction Features	Classroom Study	50 Minutes	8 Weeks	10-fold Student level	Observations
Ocumpaugh, et al. (2014) [44]	Unknown	Concentration, Confusion, Frustration Boredom	ASSISTments	Interaction Features, Student Performance	Classroom Study	~ 1 Hour	1 Day	5-fold Student level	Observations
Rodrigo and Baker (2009) [51]	40	Frustration	BlueJ Programming Environment	Interaction Features	Classroom	50 Minutes	9 Weeks	10-fold Student Level	Observations
Sabourin, Mott, and Lester (2011) [53]	450	Anxiety, Boredom, Confusion, Curiosity, Excitement, Focus, Frustration,	Crystal Island, a narrative-centered learning environment	Personal Attributes, Interaction Features	Classroom	55 minutes	1 Day	10-fold Student level	Self-report Survey
Salmeron-Majadas, Baker, Santos, Boticario (2018) [54]	41	Valence, Arousal	English Second Language Writing Task	Keyboard and Mouse Logging	Lab Study	~ 1 Hour	1 Day	10-fold Student level	Self-report Survey
Wang, Y., Heffernan, N., Heffernan, C. (2015) [60]	Unknown	Concentration, Confusion, Frustration and Boredom	ASSISTments	Interaction Features, Student Performance	Classroom Study	~ 1 Hour	1 Day	5-fold Student level	Observations

4 ONLINE LEARNING PLATFORM

Algebra Nation (AN) is a large-scale online math learning platform developed by Study Edge, an educational software and tutoring company.

Students can access AN through the website (https://www.algebranation.com/) or a mobile app. Over 150,000 students use AN each semester in Algebra 1, Geometry, and Algebra 2. Each domain uses the same interaction framework and range of activities (discussed below). Content for each domain is aligned with state education standards.

AN provides video lectures by multiple human tutors (shown in Figure 1). Each tutor provides a unique perspective on the topic, with different levels of expressiveness and technical detail. Students can watch the introductory and biographic videos of the tutors and select their favorite. AN also offers a "Test Yourself! Practice Tool", which delivers a random set of 10 quiz questions on the selected topic. These are selected from a pool of questions aligned with state math standards. Students answer questions through open text boxes, which are then automatically graded. Students may review their performance on the quiz questions, view solution videos, and revisit quiz questions and topic videos.

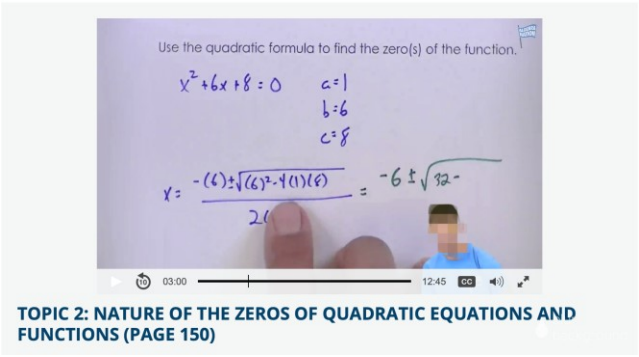

Figure 1. Sample Algebra Nation video lecture.

The AN community interacts through a *Discussion Wall* (separately for each math subject). Students post requests for help by submitting text and images, which may be answered by other students or study experts (teachers working for AN). Students who provide helpful guidance are awarded *karma points* by study experts. Karma rankings are shown in a Leaderboard and monthly prizes are given to students with the highest karma.

A suggested topic sequence is provided; however, students are free to interact with AN in whatever way they (or their teacher) choose. From reviewing the interaction data, we note that students primarily used the video lectures and the practice tests sections of AN, with the social functions being secondary.

5 DATA COLLECTION

5.1 Participants

Our collaborator (the company who runs the online platform) randomly selected 114,210 students to participate in the present study. All students attended K-12 in a state in the East Coast United States and were currently studying Algebra 1 during the 2017-2018 academic year. Students were able to opt-out of the study by ignoring our survey prompts (see below), which resulted in a 60% participation rate (N = 69,174). Study protocols were approved by our Institutional Review Board.

5.2 Affect Surveys

Affect detectors rely on "ground truth" labels for supervised classification. These labels can be obtained by acting out specific affective states, inducing affective states, or based on naturalistic affective experiences (see [27, 35, 50]). Our interest was in naturalistic affective experiences, which are provided by students themselves via self-reports or by informant reports, such as humans performing live observations or trained raters coding video data. There is no perfect method to obtain affective labels, as extensively discussed in previous publications [26, 30, 50]. Given that it was impossible to deploy human observers at the present scale and that video coding was not possible since there was no video, we utilized self-reports for affective labels.

Accordingly, in conjunction with our collaborators, we developed a mechanism to self-deliver surveys based on student activity. At each student action (e.g., selecting a video; viewing the leaderboard), a random number was generated and compared to a probability of triggering a survey prompt at that action. These probabilities were manually assigned to 19 actions and refined through a pilot process of trial and error over a period of a few weeks. Our main focus was to balance the number of surveys prompted by video viewing and practice quiz actions, which were the two most prominent student actions.

Survey prompts were displayed through a pop-up window (Figure 2), each targeting one randomly selected affective state (see below). Students could choose to close the prompt without answering the survey. Whether a student answered the survey or not, they were removed from the survey pool for two weeks. These dual requirements of easy opt-out of surveys along with infrequent prompting provided a light touch approach. The finalized survey prompts were run from September 24th

2017 to June 6th 2018 (cutoff date chosen to encompass the end of the spring semester for the majority of the school districts.). Of the 69,174 students who responded to a survey, the average number of responses per student was 1.94 (median = 1) and the maximum number of responses by any one student was 14.

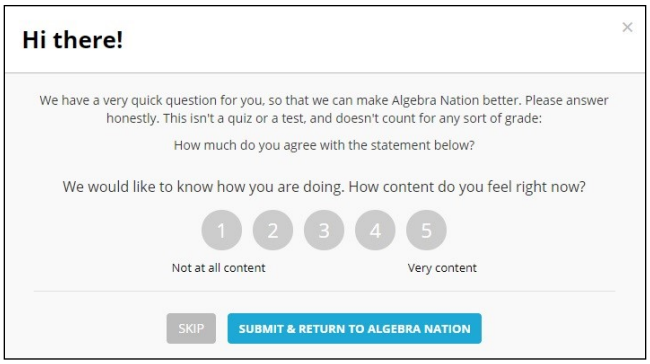

Figure 2. Example survey question given to a student while interacting with Algebra Nation.

The surveys targeted 18 affective states. These included predominant learning-centered affective states [49], such as boredom, confusion, engagement, frustration, and mind wandering (related to engagement and boredom). We also include core affective dimensions of valence and arousal [37], described here as pleasantness and wakefulness. Following best practices in survey design [4], we used a 5-point scale from "Not at all ___" to "Very ____". For unipolar prompts (e.g., not confused to very confused). Whereas 15 affective states used unipolar prompts, three states represented polar opposites (mind wandering/focused, pleasant/unpleasant, awake/sleepy), so we used a 7-point scale with contrasting options (e.g., wakefulness had a scale from "very sleepy" to "very awake") for these states. A full list of the survey prompts is shown in Table 2.

We avoided ordering effects by selecting an affective state to survey at random. This (intentionally) resulted in approximately double the number of instances for the bipolar states (see Table 3) of mind wandering, pleasantness, and arousal, where there were two questions per state. As the questions were the reverse of each other, we combined the responses of the paired questions via reverse coding. For example, the pleasantness responses were left as 1 to 7 whereas unpleasantness responses were reversed as $|x - 8|$, where x is the original survey response.

Table 2. Survey questions used for each affective state.

Affective state	Survey question
Anxiety	How anxious do you feel right now?
Boredom	How bored are you right now?
Confusion	How confused are you right now?
Contentment	How content do you feel right now?
Curiosity	How curious are you right now?
Disappointment	How disappointed do you feel right now?
Engagement	How engaged are you right now?
Frustration	How frustrated do you feel right now?
Happiness	How happy do you feel right now?
Hopefulness	How hopeful are you right now?
Interest	How interested are you right now?
Pride	How proud are you right now?
Relief	How relieved are you right now?
Sadness	How sad do you feel right now?
Surprise	How surprised are you right now?
Mind Wandering	A moment ago, my thoughts were… completely focused on other things
Mind Wandering (reverse)	…completely focused on what I was learning
Pleasantness	How pleasant do you feel right now?
Pleasantness (reverse)	How unpleasant do you feel right now?
Wakefulness	How awake are you right now?
Wakefulness (reverse)	How sleepy are you right now?

5.3 Generic Activity Features

We aimed to create student activity features that did not rely on domain-specific content (e.g., a video lecture on factoring polynomials), quiz items (e.g., solving a system of equations), or student input (e.g., reading a request for help on the discussion wall). The activity features represented counts for each action first computed in 30s chunks and then aggregated across 1, 3, and 5 minute window lengths. We varied window length to study how much data prior to a survey prompt was required for accurate affect detection. For some actions, particularly video viewing, the database sometimes recorded too many actions within 30 seconds (e.g., pausing a video hundreds of times during an interval). Although these outliers were rare, we removed them by clipping each action to a 10 count maximum per 30-second interval.

In addition to these action counts, we also computed additional features on practice quizzes. Two practice quiz features recorded whether a quiz was attempted, and whether the student returned to a previous quiz question. Our final set of 22 activity features included video viewing, practice quizzes, and discussion wall viewing, along with karma awards (based on helpful conduct in the discussion

wall) and visiting the karma Leaderboard. For a full list of features see Table 6.

Table 3. Descriptive statistics for each affect survey item.

Survey Question	N	Mean	SD	Min	Max
Anxiety	6,358	3.12	1.47	1	5
Arousal	14,843	3.64	1.94	1	7
Boredom	6,922	3.50	1.41	1	5
Confusion	6,570	2.89	1.42	1	5
Contentment	6,041	3.22	1.41	1	5
Curiosity	5,969	2.88	1.46	1	5
Disappointment	6,356	2.69	1.51	1	5
Engagement	6,269	3.20	1.39	1	5
Frustration	6,796	2.98	1.51	1	5
Happiness	6,397	3.07	1.47	1	5
Hopefulness	6,188	3.23	1.44	1	5
Interest	6,135	2.90	1.45	1	5
Mind Wandering	11,842	3.62	1.91	1	7
Pleasantness	12,398	4.13	2.11	1	7
Pride	6,265	3.04	1.44	1	5
Relief	6,181	2.93	1.46	1	5
Sadness	6,515	2.91	1.57	1	5
Surprise	5,921	2.82	1.46	1	5

6 MACHINE LEARNING[1]

We experimented with standard classifiers, feed forward and recurrent neural networks, and a genetic algorithm that learned the structure of neural networks. We also generated a chance baseline for each affective state by shuffling the survey responses and comparing to ground truth. This provided a random baseline that preserved the original distribution of responses.

6.1 Standard Classifiers

We used the scikit-learn library [47] to implement four commonly-used classifiers. These were Bayesian ridge regression, decision tree (CART [15]), Gaussian naïve Bayes and random forest. Hyperparameters for the random forest classifier and decision trees [16, 33], were tuned on the training set using the cross-validated grid search method provided by scikit-learn [47].

6.2 Neural Network Modelling

We constructed neural networks using the Keras toolkit with TensorFlow [1]. We used two model structures: feed-forward and recurrent. Our recurrent neural network contained a single long short-term memory (LSTM) activation layer. The LSTM layer implements mechanisms of forgetting and retaining information across long input sequences. LSTM models were trained on activity features computed across each 30-second time intervals (e.g., 10 sequences for the 5-min window). We also explored using bidirectional LSTMs (BLSTM), which train both forwards and backwards on these sequences. The feed-forward neural network (FFNN) used a single fully-connected activation layer (i.e., dense hidden layer). FFNN models were trained on activity across the 30-second time intervals, with no sequential information. The models used leaky rectified linear units (Leaky ReLU) as the activation function, which enabled computationally efficient training, while preventing the "dying ReLU" problem which can result in stopped training of portions of a neural network. We also used batch normalization, which regularizes activations to further increase training efficiency. Input features were normalized to the [0,1] range prior to training.

We used 32 neurons in the activation layer of all models. The model weight updates were guided by the Adam optimizer enhanced with Nesterov momentum (Nadam), which changes the magnitude of training updates on the fly to reduce the need for fine-tuning learning rates. Models were trained for 250 epochs (complete training runs through the dataset).

6.3 Genetic Algorithm

We also considered a genetic Algorithm (GA) approach – the NeuroEvolution of Augmenting Topologies (NEAT) algorithm – to evolve the topology of a neural network alongside an evolution of the network weights [59]. Because NEAT evolves both the weights and topology of the network, it must implement the genetic operators of mutation and crossover in a unique way to handle differences between network topologies. NEAT uses population speciation to track individuals with similar topologies, restricting crossover to individuals with similar network structures to ensure the resulting new topology is coherent. Mutation of the topology occurs in two ways, either by the creation of a hidden node or the addition/removal of a link between nodes. As the size of the networks may grow larger in each new generation, constraints are imposed to penalize large networks that exceed a complexity threshold.

To encourage innovation in new generations, NEAT implements speciation by grouping networks that share similar topologies into the same population. The populations are determined by a distance metric that computes the distance of a topology of an individual from the initial topology of the species. New populations are created as new networks which are dissimilar from any existing population evolved. This strategy allows the

[1]Code for models available at: github.com/emotive-computing/Hutt_CHI2019.

generation of new individuals by applying genetic operators on similar individuals in order to maintain viable network topologies without hindering the ability of the GA to develop new and unique networks.

6.4 Cross Validation

We trained our models via 10-fold student-independent cross-validation, performed separately for each survey question. Within each iteration, data from 60% of students (6 folds) were used as the training set, 30% as the validation set (to tune hyperparameters) where appropriate (see below), and 10% as the test set. Ensuring that instances from the same student are *either* in the training or testing set and inclusion of the separate validation fold increases the likelihood of model generalizability to new students.

7 RESULTS

We compare model accuracy by computing the correlation between the model predictions and the self-report survey responses. We used the Spearmen correlation coefficient (i.e., Spearman rho) since the true labels are on a Likert scale and the model predictions are continuous. All results reported are from the test folds.

7.1 Model Predictions

We examined activity periods 1, 3 and 5 minutes prior to each survey. Results were equitable across time windows when averaged across model type and affective states (see Table 4). We also examined larger time windows (7, 9 and 11 minutes) but this did not result in improved accuracy (not shown here).

Table 4. Grand mean Spearman correlation by time window.

Time Window	M	SD
1-minute	.18	.08
3-minute	.18	.09
5-minute	.19	.08
Chance	*.01*	*.02*

Figure 3 shows the mean correlations for each of the classification methods after averaging across window size and affective states. We note that the NEAT algorithm, which evolves a network along with its weights, outperformed the others on average, with Bayesian ridge regression coming a close second. NEAT evolved networks with 8 to 37 hidden nodes, suggesting considerable variability across affective states. Interestingly, the recurrent neural networks (LSTM and BLSTM) were less accurate on these data, suggesting that the sequential information was of less utility here.

Table 5 shows the spearman correlation of the best-performing model for each affective state and window size. On average, the models achieved a correlation of *rho* = 0.25 (min = 0.08, max = 0.34), which greatly exceeds the chance baseline (average of 0.01, min = -0.01, max = 0.02). The correlations were statistically significant for all affective states (using the strict threshold of $p < .001$ to account for the large number of instances) except for surprise and curiosity (ps > 0.1). The best results were achieved for confusion, frustration happiness, and hopefulness (*rho* > 0.3) and 12/18 of the correlations were higher than 0.2. These correlations, though modest, show that generic features are sufficient to predict a majority of the affective states we considered.

Table 5. Best Spearman correlation result across classifier and window for each affective state.

State	Best Result	Classifier	Window
Happiness	0.34*	NEAT	5 Min
Frustration	0.33*	NEAT	5 Min
Confusion	0.32*	NEAT	5 Min
Hopefulness	0.32*	NEAT	5 Min
Contentment	0.29*	Bayesian Ridge	5 Min
Disappointment	0.29*	NEAT	5 Min
Relief	0.29*	Bayesian Ridge	5 Min
Pride	0.28*	Bayesian Ridge	3 Min
Pleasantness	0.26*	Bayesian Ridge	5 Min
Anxiety	0.23*	NEAT	5 Min
Engagement	0.23*	Feed Forward	5 Min
Interest	0.22*	Random Forest	3 Min
Sadness	0.18*	Bayesian Ridge	3 Min
Mind Wandering	0.17*	NEAT	5 Min
Boredom	0.16*	NEAT	5 Min
Arousal	0.14*	NEAT	5 Min
Curiosity	0.10	Bayesian Ridge	5 Min
Surprise	0.08	Bayesian Ridge	5 Min
Average	0.25		

* indicates significant correlation in every fold, $p < 0.001$

7.2 Predictive Features

To further examine how individual interaction features predict affective state, we computed linear regressions for each affective state and examined the direction of the significant coefficients Table 6 ($p<0.05$ using the Benjamini & Hochberg [8] adjustment for multiple comparisons).

We note that each of the affective states, (with the exception of frustration, confusion and boredom) had a unique signature of features. For example, being awarded karma points or making a wall post were positive indicators

of arousal whereas leaving the Test Yourself suite was a negative indicator of arousal. In contrast, frustration and confusion shared the same significant features, ostensibly because they are known to co-occur [13].

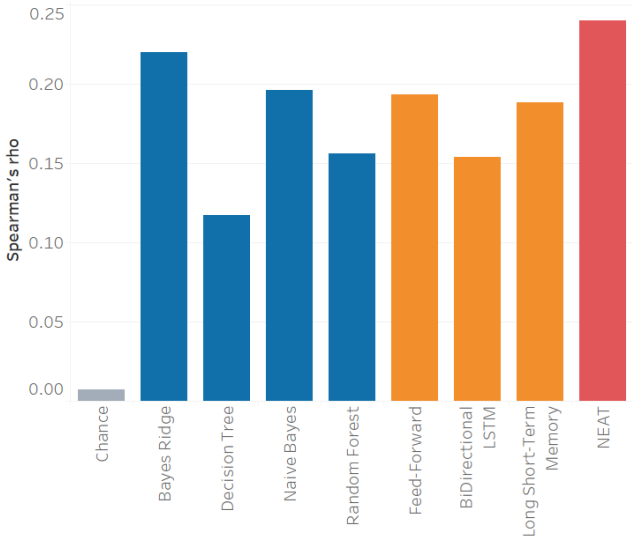

Figure 3. Model grand mean Spearman Correlation

The most predictive features across all 18 affective states were completing a video, leaving the test yourself environment, making a wall post, and reviewing the solution for a test yourself question. As one might expect, these features had a different impact depending on the affective state. Students were more content, happy, pleased, and relieved after completing a video, likely because they felt a sense of accomplishment. Conversely, they were confused, frustrated, sad, and less aroused, happy, hopeful, pleasant, proud, and relieved after exiting the Test Yourself environment. Reviewing a Test Yourself solution was associated with surprise, confusion, frustration, and disappointment, ostensibly because performance did not match expectations. Social activities, like a wall post, were associated with increased arousal, curiosity, engagement, interest, and surprise. Thus, a different profile of affective states emerged as a function of video viewing (mild positive), testing oneself (mainly negative), and engaging in social activities (positive activating).

7.3 Generalizability to a new Domain

We also collected data while a different set of students used Geometry I with Algebra Nation over the same time period. This data was used to address the question of how the models generalize to a new domain. The Geometry dataset contained 51,425 affect surveys responses (mean 2,424

Table 6. Significant features by affective state.
+ indicates a positive predictor, - indicates a negative predictor

	Anxiety	Arousal	Boredom	Confusion	Contentment	Curiosity	Disappointment	Engagement	Frustration	Happiness	Hopefulness	Interest	Mind Wandering	Pleasantness	Pride	Relief	Sadness	Surprise
Biographical Video																		
Karma Awarded		+																
Leaderboard Load																		
Profile Update																+		
Test Yourself Review Correct																		
Test Yourself Review Topic															-			+
Test Yourself Answer																		
Test Yourself Finish	+														-			
Test Yourself Load																		
Test Yourself Previous																		
Test Yourself Review Incorrect																		
Test Yourself Review Solution				+			+		+							-		+
Test Yourself Unload		-		+					+	-	-		+	-	-	-	+	
Video Caption																		
Video Completed	-			-	+		-		-	+				+		+		
Video Pause																		
Video Play																		
Video Seek																		
Video Watch																		
Wall Load																		
Wall Load More																		
Wall Post		+				+		+				+						+

responses per affective state; roughly double for bipolar states) from 28,458 students. We trained the models on one of the data sets and tested in on another – i.e., train on Algebra and test on Geometry and vice versa. The model building procedures were identical as above, except that we only focused on Bayesian ridge regression as it was more computationally efficient and generally resulted in equitable performance compared to NEAT (See Figure 3) .

The results shown in Table 7 are averaged across affective state. We note that the models generalized in both directions (Algebra 1 to Geometry and vice versa) across the three time windows. In all cases there were minimal accuracy differences (< 0.02) between training on the same domain and testing on a different domain. For example, we achieved similar average correlations when Geometry students were tested on a model trained using Algebra 1 students (*rho* = .159) or on different Geometry 1 students (*rho* = .147).

Table 7. Spearman correlations for generalizability across domains.

		Test Domain	
Time before Survey	**Train Domain**	**Algebra1**	**Geometry**
1 minute	Algebra1	0.210	0.154
	Geometry	0.188	0.145
3 minute	Algebra1	0.213	0.159
	Geometry	0.193	0.147
5 minute	Algebra1	0.213	0.159
	Geometry	0.194	0.147

8 DISCUSSION

Online learning environments provide a paradox: they greatly promote access (a positive) at the cost of meaningful engagement (a negative). An affect-sensitive online learning environment can alleviate this challenge by simultaneously attending to students' cognitive, affective, and motivational states, which comprise the three key components of learning [58]. Affect detection is a critical component of such a system. Despite considerable progress over the past two decades [17, 25], much of the work has focused on sensor-based approaches, which are not well suited for online learning. There is corresponding work on interaction-based (or sensor-free) affect detection, but this has yet to be tested at any meaningful scale – a challenge we address here. In the remainder of this section, we discuss our main findings, consider potential applications, and discuss limitations and future work.

8.1 Main Findings

Using only activity features on a large-scale dataset collected from 69,174 students over a school year, we demonstrated the feasibility of interaction-based detection of 18 affective states. Focusing on a large set of affective states enables us to derive a more complete student model which can address individual differences (e.g. less frustration for students with higher prior knowledge), and temporal changes (e.g., more anxiety as the final exam approaches) in affect.

We were able to develop affect detectors for all 18 states by using both standard and advanced machine learning methods. All detectors substantially outperformed a chance baseline and 16 out of the 18 yielded significant correlations with self-report surveys. We also found that correlations varied across the states with the highest scores for happiness, frustration, confusion and hopefulness (Spearman *rho*'s of .34, .33, .32, and .32, respectively), and the lowest scores for curiosity and surprise (*rho*'s of .10, .08, respectively). This suggests that some states can be more easily inferred from activity data than others. Further investigation is required to ascertain if an alternate feature set may produce better results for the underperforming states.

We acknowledge that these results are modest. However, sensor-free affect detection has not previously been done with such a heterogeneous sample at this scale, providing no point of comparison. Although previous research may have obtained higher accuracies, they are limited by much smaller and homogenous samples. Further, we selected a small subset of generic activity features that are more likely to generalize rather than overspecifying features to a given domain. Thus, we face a tradeoff. We improve the ecological validity and generalizability using our approach, however, a large heterogeneous sample provides more variability and a generic feature set risks underfitting, both resulting in lower accuracy.

That said, our average rho value of .25 (equivalent to a *Cohen's d* of .51) is consistent with a medium sized effect [19]. We also calculated Pearson correlations for each of our detectors, yielding an average *r* value of .22 (equivalent to a *Cohen's d* of .45), again consistent with a medium sized effect. Several of our results are within the range of previous reported research on sensor-free estimation of mental states. For example, [56], reported a correlation of .38 for detecting depression using a rich source of social media data in a study of 28,749 Facebook users. As a comparison, we obtained correlations > .30 for happiness (.34), frustration (.33), confusion (.32), and hopefulness (.32).

Similarly, [46], trained models on 66,732 Facebook users to predict five personality traits. These detectors yielded a mean correlation of .33 when evaluated on a test set of 4,824 users, again using a dataset that is far more content rich than the interaction features used in our work. Thus, though admittedly modest, the present results provide a useful baseline for what can be achieved with a set of generic activity features are used for affect detection in a large, heterogeneous dataset.

In terms of generalizability, our results are consistent with previous interaction-based affect detectors that generalize to new students. However, previous research has largely used features tailored to specific domains and learning platforms, which makes generalizability across domains and platforms implausible. By mainly considering generic features, such as video viewing behavior and forum posts, we showed that our models generalized to a second domain albeit with the same learning environment. In a related vein, whereas previous work restricted data collection to a lab or classroom context, we had no control on the learning context (home, school, afterschool, while commuting) and how students chose to use the system, suggesting that the models also likely generalize to multiple contexts (though we have not empirically shown this yet).

Finally, we investigated patterns between activity features and affective states. We discovered that with the exception of boredom, the other 17 affective states had at least one single feature as a significant predictor. Importantly, of the 22 features considered, only eight were predictive of at least one affective states. These eight actions, associated with videos, Test Yourself items, and Wall posts, can be linked to unique learning functions/phases with distinct affective profiles. Specifically, video completion is associated with information acquisition, test yourself with information retrieval, and Wall posting with social functions.

8.2 Applications

The key application of this work is to integrate the affect detectors into Algebra Nation, so that the system may detect affective states in real-time. The resultant data can be used in a number of ways, beginning with better understanding students' affective experience during learning with this platform. Are certain videos particularly engaging, inspiring, or motivating? Do others make students' minds wander or induce boredom? Do some test items inspire a sense of accomplishment and hope whereas others lead to disappointment and despair? Is the wording of some items simply too confusing and in need of rephrasing? Similarly, when testing new content, frustrating questions or boring videos could be identified at an early stage and revised before integration into the system. In addition to informing instructional design, automated affective reports can be provided to teachers (in anonymized and aggregate form), so they can adapt their pedagogical approach as well.

Affect detection also presents the possibility to develop interventions to help students regulate their affect. For example, students who show signs of frustration with a certain topic might be referred to a particular video that may help them. In contrast, students who are engaged while viewing a video may be encouraged to try a quiz on that topic. Research is needed to identify the optimal contextually-grounded strategies for the different affective states. It is also important to note that affect detection is inherently imperfect. The system might detect frustration when the student is actually content. However, affect detection does not need to be perfect as long as we account for its imperfection when designing intervention strategies. For example, Algebra Nation could take a probabilistic approach to delivering interventions where the detector's confidence determines when and which interventions are launched. Similar accuracies have been used to trigger successful interventions in the past [22] but importantly, interventions should be constructed to be fail-soft in that there are no harmful effects if delivered incorrectly. For example, intervening infrequently and allowing students to choose if they want to engage in the intervention are some possible ways to accomplish this.

Finally, measurement is a critical component of science, and affect detection at scale can advance basic research on affect and learning. Our current understanding of how affect interacts with cognition to influence learning is based on rigorous scientific research, but on small samples and timescales ([6, 7, 17, 44]). Automated affect sensing at scale and across time can both complement traditional research by testing existing theories while also advancing basic research by discovering new insights. For example, foundational question of how individual differences and contextual factors interact to influence and affect and cognition have been left unanswered due to small and mostly homogenous research samples. Automated affect detection on large heterogeneous samples can provide a critical piece of the puzzle.

8.3 Limitations

Like all studies, ours has limitations. As with any complex psychological constructs, there is no "direct" way to

measure affect [52], so one has to rely on operational definitions of the construct. We chose self-reports collected via an experience sampling method as our operational definition of affect, due to multiple constraints articulated in the Introduction. This choice has strengths and limitations as discussed extensively in [26, 30, 50]. As such, all conclusion we draw from this research are restricted to our specific conceptualization of affect. Relatedly, our light-touch survey approach, which only measures one affective state per survey, ignores the potential of co-occurrence affective states or that multiple states could be occurring in the same five minute window.

Second, though a strength of this work was the use of generic features, this also presents a limitation. A generic feature set operates at a higher level of abstraction, which may aid generalizability at the cost of accuracy. Indeed, our average correlation of 0.25 is consistent with a medium sized effect [19], suggesting there is considerable room for improvement. It would be interesting to further explore the accuracy/generalizability tradeoff by contrasting with low-level content-specific features, such as particular videos viewed and specific questions attempted.

Finally, we only considered mathematics topics. Although we have shown that the models generalize across two mathematics domains, it is unclear how they would perform on other topics, such as foreign languages. Similarly, all data was collected from users from a single U.S. state using one learning platform, so it is unclear how the models will generalize to students from other states using similar online learning technologies.

8.4 Future Work

In addition to addressing the limitations described above, there are also several promising avenues for future work. First, data collection is still ongoing, allowing us to explore how well these models generalize to a new academic year. There is also the potential to explore how these models generalize to students in a different state as Algebra Nation is being expanded across the nation.

Second, we will investigate how the models generalize to different topics, either other mathematics topics within Algebra Nation, (e.g. Algebra 2) or additional topics in similar learning environments. It will be particularly informative to discover which activity features generalize across environments and whether the links between activity features and affective states replicate.

Third, we will consider ways to enhance our detection models. One idea is to build multiple models per affective state and combine predictions via ensemble approaches. A further idea would be to harness user characteristics in our model as done in previous work [7]. By examining how usage patterns can be used to group users, we can personalize models for each sub-group and even refine them using active learning methods [10].

Finally, we are interested to see how the affective states, measured via self-reports and our automated detectors, predict critical educational outcomes, such as end of year tests. This information will be essential to design affect-sensitive interventions that help to improve learning outcomes by responding to affect.

9 CONCLUDING REMARKS

Online learning environments present new opportunities and new challenges to students and educators alike. Understanding students' affective experience is one important way to address some of the challenges with these environments. Our results that we can model students' affective states from their digital traces with these learning environments at a previously unexplored scale across an extended time frame and in a generalizable fashion. In doing so, we have advanced the field of affect computing by scaling up affect detection.

ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305C160004 and Intel Research. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education

REFERENCES

[1] Abadi, M. et al. 2015. TensorFlow: large-scale machine learning on heterogeneous systems.

[2] Ai, H. et al. 2006. Using system and user performance features to improve emotion detection in spoken tutoring dialogs. *Interspeech* (2006).

[3] Algebra Nation: 2018. *https://www.algebranation.com/fl/*.

[4] Allen, I.E. and Seaman, C.A. 2007. Likert scales and data analyses. *Quality progress*. 40, 7 (2007), 64–65.

[5] Arroyo, I. et al. 2009. Emotion sensors go to school. *Frontiers in Artificial Intelligence and Applications* (2009), 17–24.

[6] Baker, R.S.J. d. et al. 2012. Towards sensor-free affect detection in cognitive tutor algebra. *Proceedings of the 5th International Conference on Educational Data Mining*. (2012).

[7] Baker, R.S.J. d. and Ocumpaugh, J. 2015. Interaction-based affect detection in educational software. *The Oxford Handbook of Affective Computing*. R. Calvo et al., eds.

[8] Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple controlling the false discovery rate: a practical and powerful approach to multiple

testing. *Journal of the Royal Statistical Society*. 57, 1 (1995), 289–300. DOI:https://doi.org/10.2307/2346101.

[9] Boekaerts, M. and Pekrun, R. 2016. Emotions and emotion regulation in academic settings. *Handbook of educational psychology*. 76–90.

[10] Bonwell, C.C. and Eison, J.A. 1991. *Active learning: Creating excitement in the classroom. ERIC digest.*

[11] Bosch, N. et al. 2015. Automatic detection of learning-centered affective states in the wild. *Proceedings of the 20th International Conference on Intelligent User Interfaces* (New York, NY, USA, 2015), 379–388.

[12] Bosch, N. et al. 2016. Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems*. 6, 2 (2016), 1–26. DOI:https://doi.org/10.1145/2946837.

[13] Bosch, N. and D'Mello, S. 2017. The affective experience of novice computer programmers. *International Journal of Artificial Intelligence in Education*. 27, 1 (2017), 181–206. DOI:https://doi.org/10.1007/s40593-015-0069-5.

[14] Botelho, A.F. et al. 2017. Improving sensor-free affect detection using deep learning. *Artificial Intelligence in Education*.

[15] Breiman, L. et al. 1984. *Classification and regression trees*.

[16] Breiman, L. 2001. Random forests. *Machine Learning*. 45, 1 (2001), 5–32. DOI:https://doi.org/10.1023/A:1010933404324.

[17] Calvo, R.A. and D'Mello, S.K. 2010. Affect detection: an interdisciplinary review of models, methods, and their applications. *IEEE Trans. Affect. Comput.* 1, 1 (2010), 18–37. DOI:https://doi.org/10.1109/T-AFFC.2010.1.

[18] Clow, D. 2013. MOOCs and the funnel of participation. *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK '13* (2013), 185.

[19] Cohen, J. 2013. *Statistical power analysis for the behavioral sciences*. Taylor & Francis.

[20] Conati, C. 2002. Probabilistic assessment of user's emotions in educational games. *Applied Artificial Intelligence*. 16, 7–8 (2002), 555–575. DOI:https://doi.org/10.1080/08839510290030390.

[21] Conati, C. and MacLaren, H. 2009. Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*. (2009). DOI:https://doi.org/10.1007/s11257-009-9062-8.

[22] D'Mello, S.K. et al. 2016. Attending to attention: detecting and combating mind wandering during computerized reading. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (2016), 1661–1669.

[23] D'Mello, S.K. et al. 2008. Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction*. 18, 1–2 (2008), 45–80. DOI:https://doi.org/10.1007/s11257-007-9037-6.

[24] D'Mello, S.K. et al. 2005. Integrating affect sensors in an intelligent tutoring system. *In Affective Interactions: The Computer in the Affective Loop Workshop at 2005 International Conference on Intelligent User Interfaces*. (2005), 7–13. DOI:https://doi.org/10.1158/0008-5472.CAN-06-1527.

[25] D'Mello, S.K. et al. 2018. Multimodal-multisensor affect detection. *The Handbook of Multimodal-Multisensor Interfaces*. S. Oviatt et al., eds. ACM Books/Morgan Claypool. 167–202.

[26] D'Mello, S.K. 2016. On the influence of an iterative affect annotation approach on inter-observer and self-observer reliability. *IEEE Transactions on Affective Computing*. 7, 2 (2016), 136–149. DOI:https://doi.org/10.1109/TAFFC.2015.2457413.

[27] D'Mello, S.K. et al. 2018. The affective computing approach to affect measurement. *Emotion Review*. 10, 2 (2018), 174–183. DOI:https://doi.org/10.1177/1754073917696583.

[28] D'Mello, S.K. and Kory, J. 2012. Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. *ACM International Conference on Multimodal Interaction*. (2012). DOI:https://doi.org/10.1145/2388676.2388686.

[29] DeFalco, J.A. et al. 2018. Detecting and addressing frustration in a serious game for military training. *International Journal of Artificial Intelligence in Education*. 28, 2 (Jun. 2018), 152–193. DOI:https://doi.org/10.1007/s40593-017-0152-1.

[30] Douglas-Cowie, E. et al. 2005. Multimodal databases of everyday emotion: facing up to complexity. *Interspeech 2005*. (2005), 813–816.

[31] Goleman, D. 1995. *Emotional intelligence*.

[32] Grafsgaard, J.F. et al. 2013. Automatically recognizing facial indicators of frustration: a learning-centric analysis. *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (Washington, DC, USA, 2013), 159–165.

[33] Ho, T.K. 1995. Random decision forests. *Proceedings of the Third International Conference on Document Analysis and Recognition* (Washington, DC, USA, 1995), 278–282.

[34] Hutt, S. et al. 2017. "Out of the fr-eye-ing pan": towards gaze-based models of attention during learning with technology in the classroom. *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (New York, NY, USA, 2017), 94–103.

[35] Krosnick, J.A. 1999. Survey research. *Annual review of psychology*.

[36] Lee, D.M.C. et al. 2011. Exploring the relationship between novice programmer confusion and achievement. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2011).

[37] Linnenbrink, E.A. 2007. The role of affect in student learning: a multi-dimensional approach to considering the interaction of affect, motivation, and engagement. *Emotion in Education*. R. Pekrun, ed. Elsevier. 107–124.

[38] Liyanagunawardena, T.R. et al. 2013. {MOOCs}: {a} systematic study of the published literature 2008-2012. *The International Review of Research in Open and Distributed Learning*. 14, 3 (2013), 202–227.

[39] Liyanagunawardena, T.R. 2013. MOOCs: a systematic study of the published literature 2008-2012. *International Review of Research in Open & Distance Learning*. 14, (2013), 202–227.

[40] McDaniel, B. et al. 2007. Facial features for affective state detection in learning environments. *Proceedings of the Annual Meeting of the Cognitive Science Society* (Jan. 2007).

[41] Mcinnerney, J.M. and Roberts, T.S. 2004. Online learning: social interaction and the creation of a sense of community what is isolation? *Sociology The Journal Of The British Sociological Association*. 7, 3 (2004), 73–81. DOI:https://doi.org/10.1.1.99.9614.

[42] Mota, S. and Picard, R.W. 2003. Automated posture analysis for detecting learner's interest level. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (2003).

[43] My Maths: 2018. *https://www.mymaths.co.uk/secondary.html*.

[44] Ocumpaugh, J. et al. 2014. Population validity for educational data mining models: a case study in affect detection. *British Journal of Educational Technology*. (2014). DOI:https://doi.org/10.1111/bjet.12156.

[45] Pardos, Z.A. et al. 2014. Affective states and state tests: investigating how affect throughout the school year predicts end of year learning outcomes. *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK '13*. 1, (2014), 117–124. DOI:https://doi.org/10.1145/2460296.2460320.

[46] Park, G. et al. 2015. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*. (2015). DOI:https://doi.org/10.1037/pspp0000020.

[47] Pedregosa, F. et al. 2011. Scikit-learn: machine learning in python. *Journal of Machine Learning Research*. 12, (2011), 2825–2830.

[48] Pekrun, R. 2017. Emotion and achievement during adolescence. *Child Development Perspectives*. 11, 3 (2017), 215–221. DOI:https://doi.org/10.1111/cdep.12237.

[49] Pekrun, R. 2007. Emotions in students' scholastic development. *The scholarship of teaching and learning in higher education: An evidence-based perspective*. 553–610.

[50] Porayska-Pomsta, K. et al. 2013. Knowledge elicitation methods for affect modelling in education. *International Journal of Artificial Intelligence in Education*. 22, 3 (2013), 107–140. DOI:https://doi.org/10.3233/JAI-130032.

[51] Rodrigo, M.M.T. and Baker, R.S.J. d. 2009. Coarse-grained detection of student frustration in an introductory programming course. *Proceedings of the fifth international workshop on Computing education research workshop - ICER '09*. (2009), 75. DOI:https://doi.org/10.1145/1584322.1584332.

[52] Rosenthal, R. and Rosnow, R.L. 1991. Essentials of behavioral research: methods and data analysis. *Boston, MA*. (1991).

[53] Sabourin, J. et al. 2013. Discovering behavior patterns of self-regulated learners in an inquiry-based learning environment. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2013).

[54] Salmeron-Majadas, S. et al. 2018. A machine learning approach to leverage individual keyboard and mouse interaction behavior from multiple users in real-world learning scenarios. *IEEE Access*. (2018). DOI:https://doi.org/10.1109/ACCESS.2018.2854966.

[55] Sandeen, C. 2013. Integrating moocs into traditional higher education: the emerging "mooc 3.0" era. *Change: The Magazine of Higher Learning*. 45, 6 (Nov. 2013), 34–39. DOI:https://doi.org/10.1080/00091383.2013.842103.

[56] Schwartz, H.A. et al. 2014. Towards assessing changes in degree of depression through facebook. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (2014).

[57] Smallwood, J. et al. 2008. When attention matters: the curious incident of the wandering mind. *Memory & Cognition*. 36, 6 (Sep. 2008), 1144–1150. DOI:https://doi.org/10.3758/MC.36.6.1144.

[58] Snow, R. et al. 1996. Individual differences in affective and conative functions. *Handbook of educational psychology*. 243–310.

[59] Stanley, K.O. and Miikkulainen, R. 2002. Evolving neural networks through augmenting topologies. *Evolutionary Computation*. 10, 2 (2002), 99–127.

[60] Wang, Y. et al. 2015. Towards better affect detectors: effect of missing skills, class features and common wrong answers. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK '15*. (2015). DOI:https://doi.org/10.1145/2723576.2723618.