Chapter ?: Modelling Affect by Mining Students' Interactions within Learning Environments

Manolis Mavrikis[*,1], Sidney D'Mello[2], Kaska Porayska-Pomsta[1], Mihaela Cocea[1], and Art Graesser[2]
[1]London Knowledge Lab, The University of London,
[2]Institute for Intelligent Systems, The University of Memphis

INTRODUCTION

In the past decade research on affect-sensitive learning environments has emerged as an important area in Artificial Intelligence in Education (AIEd) and Intelligent Tutoring Systems (ITS) [1-6]. These systems aspire to enhance the effectiveness of computer mediated tutorial interactions by dynamically adapting to individual learners' affective and cognitive states [7] thereby emulating accomplished human tutors [7,8]. Such dynamic adaptation requires the implementation of an *affective loop* [9] , consisting of: (1) detection of the learner's affective states, (2) selection of systems actions that are sensitive to a learner's affective and cognitive states, and sometimes (3) synthesis of emotional expressions by animated pedagogical agents that simulate human tutors or peer learning companions [9,10].

The design of affect-sensitive learning environments is grounded in research that states that the complex interplay between affect and cognition during learning activities is of crucial importance to facilitating learning of complex topics, particularly at deeper levels of comprehension [11-17]. Within this research, one particular area of interest is concerned with the ways in which human tutors detect and respond to learners' affective states: robust detection of learners' affect is critical to enabling the development of affect-sensitive ITSs. In this chapter we will examine the state-of-the art methods by

* Corresponding author: m.mavrikis@lkl.ac.uk

which such detection can be facilitated. In particular, we examine the issues that arise from the use of supervised machine learning techniques as a method for inferring learners' affective states based on features extracted from students' naturalistic interactions with computerized learning environments. Our focus is on general methodological questions related to educational data mining (EDM) with an emphasis on data collection protocols and machine learning techniques for modeling learners' affective states. We present two case studies that demonstrate how particular EDM techniques are used to detect learners' affective states based on parameters that are collected during learners' interactions with different learning environments. We conclude with a critical analysis of the specific research outcomes afforded by the methods and techniques employed in the case studies that we present.

I. BACKGROUND

There is an increasing body of research that is concerned with identifying the affective states that accompany learning and devising ways to automatically detect them during real interactions within different educational systems. [18-24].

Different approaches to detecting affect focus on monitoring facial expressions, acoustic-prosodic features of speech, gross body language, and physiological measures such as skin conductivity or heart rate monitoring,. For extensive reviews the reader is referred to [25-31]. Another approach involves an analysis of a combination of lexical and discourse features with acoustic-prosodic and lexical features obtained through a learner's interaction with spoken dialogue systems. A number of research groups have reported that appending an acoustic-prosodic and lexical feature vector with dialogue features results in a 1-4% improvement in classification accuracy [32-34]. Whilst these approaches have been shown to be relatively successful in detecting affective states of the learners,

some tend to be quite expensive and some of them can be intrusive and may interfere with the learning process.

An interesting alternative to physiological and bodily measures for affect detection is to focus on learners' actions that are observable in-the-heat-of-the-moment, i.e. at the time at which they are produced by the learner in the specific learning environment [6]. One obvious advantage of referring to these actions is that the technology required is less expensive and less intrusive than physiological sensors. Furthermore, by recording learner's actions as they occur it is possible to avoid imposing additional constraints on the learner (i.e. no cameras, gloves, head gear etc.), thereby also reducing the risk of interference with the actual learning process and it somewhat alleviates the concern that learners might disguise the expression of certain negative emotions.

In this chapter we present two case studies which differ from the previous research in that we focus on interaction logs as the primarily means to infer learner affect. The first study considers a broad set of features, including lexical, semantic, and contextual cues, as the basis for detecting learners' affective states. The second case study employs an approach that relies only on student actions on the interactive feature of the environment such as hint or information buttons to infer learner affect.

The following section discusses general methodological considerations behind these two approaches before presenting the two case studies in more detail.

II. METHODOLOGICAL CONSIDERATIONS

One of the most important considerations when attempting to detect affect from interaction features is the context in which the data are collected. Ideally, a study through which such data is collected would achieve high ecological validity, i.e. it should approximate the situation under which the results are

expected to generalise [35]. It is also desirable that studies of learners' affect would involve learners who are familiar with a specific learning environment, who can use the environment in their own time and location over an extended period of time, and who have real learning objectives in relation to the domain under investigation. While achieving ecological validity is desirable in many studies, it is essential in any investigation of human affect. Context influences affective states, hence, modelling the contextual underpinnings of learners' affective experiences is critical to obtaining valid and generalizable research results [35-38].

However, such flexibility and familiarity with both the domain and the environment is not always possible given the requirements of the data mining techniques currently available. For example supervised learning, the most widely used data mining technique, requires training based on labeled instances (also referred to as *ground truth*) in order to relate affect categories to the interaction parameters and bodily and physiological channels. Collecting these affect labels requires a compromise on ecological validity. We illustrate this issue by examining the two main methods used to collect affect labels.

The first method involves concurrent reports of affective states provided either by the learners themselves (self-report) or by external observers (e.g. tutors or peer students). For example, [39] implemented an *emote-aloud* protocol that allowed for the collection of self reports in real time while students are interacting with AutoTutor, an Intelligent Tutoring System (ITS) with conversational dialogue [40]. The emote-aloud procedure is a modification of the *think-aloud* procedure [41], which involves participants talking about their cognitive states and processes while working on tasks that require deeper cognitive engagement, such as solving problems [41] or comprehending text [42]. Emoting-aloud involves participants verbalising their states on a moment-by-moment basis while interacting with a learning environment, except that the verbalizations are of affective rather than

cognitive states and processes. A similar technique during a computer-mediated tutorial for recording tutors' annotations of learner's affect was employed in [6]. Online affect judgments by observers, such as observations of students' in a classroom environment [43], can provide an alternative to self reports.

The second method is to employ a retrospective affect judgment protocol as an offline measure of learners' affect [44] Specific techniques might involve students watching replays of their interactions with a learning environment in order to report on their affective states [e.g., 45]. A more elaborate example is provided in [39] where videos of participant's face and computer screen were synchronized and displayed to the learners after the tutoring session with AutoTutor in order to enable them to make judgments on which affective states were present at various points in the session. In addition to self-reports retrospective *post-task* annotations can involve peers and/or tutors in walkthroughs on replays of learners' interactions. For example, in [46]experienced tutors were asked to annotate replays of students' interactions with a web-based interactive learning environment, while [47] used an untrained peer and two trained judges to obtain affect labels. Alternatively, it is also possible to annotate for affect appropriate log files [24].

The methods listed above have their possible advantages and disadvantages. The reader is referred to [46,48], where these are discussed in detail. The rest of this chapter presents two case studies that demonstrate the use of data-collection methodology and data mining techniques.

III. CASE STUDIES

A. Detecting Affect from Dialogues with AutoTutor

1. Context

This case study is derived from a larger project that aspires to integrate state of the art affect sensing devices into an existing ITS called AutoTutor [40]. AutoTutor is a fully automated computer tutor that

helps students learn Newtonian physics, computer literacy, and critical thinking by presenting challenging problems (or questions) that require reasoning and explanations in the answers. AutoTutor and the learner collaboratively answer these difficult questions via a mixed-initiative dialogue that is dynamically adaptive to the cognitive states of the learner.

The AutoTutor research team is currently working on a version of AutoTutor that is sensitive to the affective as well as the cognitive states of the learner [5]. The affective states being tracked are boredom, engagement/flow, confusion, frustration, and delight. These were the most prominent affective states that were observed across multiple studies with AutoTutor and other learning environments [39,49,50].

2. Mining Dialogue Features from AutoTutor's Log Files

The data analysis described here was conducted on a corpora obtained by conducting two studies where learners' affective states and dialogue patterns were recorded during interactions with AutoTutor. The first study [39] implemented an emote-aloud protocol with seven participants. The second ($N = 28$) implemented an offline retrospective affect judgment protocol where the affect judges were the participants, an untrained peer, and two trained judges. In both studies participants were tutored on computer literacy topics (hardware, internet, and operating systems) with AutoTutor.

Several features from AutoTutor's log files were mined in order to explore the links between the dialogue features and the affective states of the learners. These features included temporal assessments for each student-tutor turn such as the *subtopic number*, the *turn number*, and the student's *reaction time* (interval between presentation of the question and the submission of the student's answer). Assessments of response verbosity included the *number of characters* (letters, numbers) and *speech act* (that is, whether the student's response was a contribution towards an answer versus a frozen expression, e.g., "I don't know, " "Uh huh"). The conceptual quality of the student's response was evaluated by Latent

Semantic Analysis (LSA) [51]. LSA is a statistical technique that measures the conceptual similarity of two text sources. LSA-based measures included a *local good score* (the conceptual similarity between the student's current response and the particular expectation, i.e. ideal answer, being covered) and a *global good score* (the similarity of set of student responses to a problem and the set of expectations in a good answer). Additionally, changes in these measures when compared to the previous turn were also included as the *delta local good score* and the *delta global good* score. AutoTutor's major dialogue moves were ordered onto a scale of conversational *directness,* ranging from -1 to 1, in terms of the amount of information the tutor explicitly provides the student. AutoTutor's short *feedback* (negative, neutral negative, neutral, neutral positive, positive) is manifested in its verbal content, intonation, and a host of other non-verbal cues. The feedback was aligned on a 5 point scale ranging from -1 (negative) to 5 (positive feedback).

3. Automated Dialogue-Based Affect Classifiers

Statistical patterns between the affective states and dialogue features are extensively discussed in previous publications [39,52]. For example, boredom occurs later in the session (high subtopic number), after multiple attempts to answer the main question (high turn number), and when AutoTutor gives more direct dialogue moves (high directness). The focus of this chapter is on the accuracy by which several standard classification algorithms could individually distinguish each affective state from neutral (no affect) as well as collectively discriminate between the affective states. The classification algorithms tested were selected from a list of categories including Bayesian classifiers (Naive Bayes and Naive Bayes Updatable), functions (Logistic Regression, Multilayer Perceptron, and Support Vector Machines), instance based techniques (Nearest Neighbor, K*, Locally Weighted Learning), meta classification schemes (AdaBoost, Bagging Predictors, Additive Logistic Regression), trees (C4.5 Decision Trees, Logistic Model Trees, REP Tree), and rules (Decision Tables, Nearest Neighbour

Generalization, PART).

Machine learning experiments indicated that classifiers were moderately successful in discriminating the affective states of boredom, confusion, flow, frustration, and neutral, yielding a peak accuracy of 42% with neutral (chance = 20%) and 54% without neutral (chance = 25%). Individual detections of boredom, confusion, flow, and frustration, when contrasted with neutral, had maximum accuracies of 69%, 68%, 71%, and 78%, respectively (chance = 50%). These results support the notion that dialogue features is a reasonable source for measuring the affective states that a learner is experiencing. Comparisons among the different types of classifiers indicated that functions, meta, and tree based classifiers were similar quantitatively and yielded significantly higher performance than the other categories. Classification accuracy scores for the instance-based learning were significantly lower than the other five classifier categories. Bayesian classifiers outperformed rule based classification schemes. In general AdaBoost, logistic regression, and C4.5 decision trees yielded the best performance.

B. Case Study 2: Predictive Modelling of Student-Reported Affect from Web-Based Interactions in WaLLiS

1. Context

This case study is derived from a research project aiming at the enhancement of an existing web-based intelligent learning environment (WaLLiS [53]) with low-cost capabilities for affect detection and affect-sensitive responses. For a brief description of WaLLiS see also Chapter x, this book.

One of the aims of the research described here was to develop a methodology particularly suited for investigating affective factors under ecologically valid situations. In response to a desire to record

data under the most realistic conditions, and because of the difficulty of conducting emote-aloud protocols at students' homes, retrospective walkthroughs with students and tutors were conducted. A representative sample of 18 students was obtained out of 209 students who were already familiar with WaLLiS. These students were using the system at their own time and location, while they were studying for a real course of their immediate interest. The sample was selected on the basis of disproportionate stratified random sampling [54] and included students with different mathematical abilities and awareness of their own abilities.

Similar to the previous case study, data was collected from two studies, one with students retrospectively reporting on their own affect and one with tutors watching replays of students' interactions. Only the first study collected substantial amount of data to conduct a machine-learning analysis. The second met significant difficulties, because, in the absence of other information (e.g. participants' face) tutors found it very difficult to provide judgments of students' affect. Hence, this study was only used for qualitative analysis and to validate the models derived from the machine learning procedures as described below.

2. Machine learned models from student-system interactions

Similarly to the Case Study 1, comparisons of different machine learning techniques were performed. The overall aim of this research was to: a) derive hypotheses for future research and b) enable triangulation of the results with a qualitative analysis of the students' and tutors' walkthroughs. For this reason decision trees were chosen as the classification method. This choice was motivated by the fact that decisions trees are relatively inspectable and afford easy comparison and consolidation of derived models.

Separate models were derived for each affective factor rather than, as is usually the case, relying

on one model that would predict all the factors. Accordingly, the machine-learning algorithm is presented with pre-processed vectors (instances) automatically constructed from the raw data by an extraction tool that matched the timestamps of contextual factors (correctness of answer, question difficulty, time spent) with the corresponding student's report. These vectors consist of the contextual factors as features and a nominal class that encodes students' reports. In the case of the affective factors the values are binary indicating presence of absence of *frustration*, *boredom*, *confusion*. For *confidence*, *interest*, and *effort* the class takes values that depict the relative change of each factor: *decrease*, *increase* and *extreme_decrease*, *extreme_increase*.

Due to the limited size of the data, the majority of the reported affective characteristics pertain to the *confidence* and *effort* factors and therefore the machine learning analysis exclusively focused on these two factors.

As mentioned above, there was an explicit attempt to take into account the history of students' actions. History is represented as a vector, the elements of which encode the number of times that each type of action (e.g. a hint) occurred in a time window. This window spans back to the last change of the factor under investigation or (if it has not recently changed) to the start of the relevant situation or exercise.

Space constraints prevent us from discussing all the results that constitute the rules and future hypothesis that were derived from the decision trees. These are discussed in detail in other publications [45,46]. Here, we use confidence as an example to depict the methodology and the type of results that were obtained.

---- INSERT FIGURE confidence.eps AROUND HERE ---

Figure 1. Graphical representation of the decision tree for confidence. Each node represents an attribute and the labels on the edges between nodes indicate the possible values of the parent attribute. Following a path from root to a leaf results in a rule that shows the value of the factor over the values of the attributes in the path. The numbers in brackets next to the designated class indicate the number of instances correctly classified by this rule over the misclassified ones. Nodes with layers (e.g. node a) originate from a *history* vector. For example AnswerIncorr > 2 shows that in the relevant time window more than two incorrect answers were provided by the student.

In total there were 289 reports in relation to confidence; these comprise set A. Running the algorithm on this set of data leads to biased rules since it contains only instances where changes in affective states are reported. It is important, however, to train the model with instances of the same patterns where there are no changes to the affective characteristics. Therefore, we extracted the instances (249 in total) where the same actions as these of set A occurred but were not associated with a particular report; these comprise set B. Figure1 shows a graphic representation of the tree resulting from the merged sets of instances with attributes *action*, *difficulty*, *student ability*, *last answer*, *last feedback*, *time*, and the *history* vector. As an example, the rules associated with the confirm answer action of students are described. The tree suggests that when students confirm an incorrect answer (node a) after having previously requested at least one hint, their confidence decreases (leaf b). If no hints were requested leaf (c) suggests that if they previously had many (more than two) incorrect answers then they

report that their confidence decreases (the two misclassified instances in leaf c are of an extreme decrease). Otherwise it seems that the outcome depends on students' previous knowledge, the difficulty of the question, and the necessity of the help request (node d).

The rest of the tree can be interpreted in a similar fashion. For example, the rule associated with node (e), where students are requesting hints after submitting a wrong answer shows that students' reports vary depending on whether or not their previous answer was partially correct. A closer inspection of the rule suggests that in situations where the students provided a partially correct answer and where the system responds to such an answer with negative feedback, the students' confidence level tends to drop. This is particularly the case for students who do not spent sufficient time to read and interpret the hints provided by the system (node f).

Overall, with the addition of history in the vector, cross-validation performed in the decision tree for confidence indicated that the tree correctly classified 90.91% of the cases (Kappa=0.87). Accuracy for effort was 89.16%, Kappa=0.79; these can be considered as satisfactory results. Although, in this case, the addition of history did not improve the results significantly, such information can be very useful in situations like open-learner modeling where it would be important to communicate to the student the rationale behind the systems' decisions.

The results demonstrate that rule induction provides a mechanism for deriving a predictive model in the form of rules that is based on students' actions with the system. Most of these rules are intuitive but defining them by hand would have required a thorough, operational understanding of the processes involved, not easily achieved by experts in the field. Although the process for collecting data was time consuming and led to small amount of unequivocal rules, the methodology and machine learning method is generalisable to different situations resulting in at least hypotheses about rules that can guide the design of future studies.

IV. DISCUSSION

The case studies presented in this chapter demonstrate the benefits of using EDM techniques to monitor complex mental states in educational settings. They also identify several important challenges for the EDM field, particularly in relation to the prediction of learner affect. We highlighted different methods for collecting and annotating data of students' interaction, the importance of ecological validity, as well as the difficulty of achieving it. Several chapters in this book provide examples of the types of analysis and research that recently has become possible owing to the availability of data from systems integrated in real pedagogical situations. However, compared to other applications of EDM, affect prediction from data introduces additional challenges to the ones faced when investigating, for example, the effects of the interaction in learning.

Although supervised learning approaches require a measure of ground truth in the phenomenon being predicted, additional challenges arise in affect measurement. We discussed the need for employing data that are representative of the behavioural evidence to which an ITS has runtime access. However, affect is a psychological construct that is notoriously difficult to measure since human judgments of affect are often vague, ill-defined, and possibly indeterminate. Self-reports, commonly used to elicit such judgments, also present a number of possible problems, such as the bias of subjectivity resulting from the fact that each affective characteristic may have a different meaning for different learners. Therefore, it is important to look at protocols that extend beyond self reports for monitoring affective states. For example in case study 1, the video recordings of learners' facial expressions and the dialogue cues enabled the judges to make more informed ratings. Furthermore, reliable physiological sensors can also be employed. Although this may involve violating the ecological validity principle, it may be necessary for deriving models that can be introduced and evaluated subsequently in more ecologically valid situations.

Another important consideration comes from the use of qualitative methods for data-collecting. These often result in limited and sparse data. Therefore, it is common to employ cross-validation to assess the quality of the predictive method. Since any evaluation influences the validity of the results, the stratified cross-validation that is typically employed in the field could benefit from being performed in a way that takes into account that we are dealing with educational data. The 'off-the-shelf' methods for cross-validation could potentially introduce biases in the evaluation by including instances of the same learner in both the training and testing sets. Although in some cases this is not a problem, in situations where the nature of the predicted class can be affected by latent inherent characteristics of the learner (such as affective traits or prolonged mood states), the evaluation can be biased.

Furthermore, instead of 'blindly' evaluating a model's performance in relation to existing data, cost-sensitive classification [55] can provide more practical results. That is, instead of assuming equality between the costs of false positives and false negatives, it may be worth taking into account the cost of misclassifications. For example, [45] describe the use of such an approach based on rules derived from experts who evaluated the effect of false positives for a model of confidence prediction in various situations. The estimated 'cost' is then employed at prediction time to inform the pedagogical decisions of the system. Having a cost function also enables the implementation of algorithms that take the cost into account during learning, thus leading to a model that is optimized with respect to the cost of the adaptive decisions taken rather than just against the data available alone.

A related question refers to the extent to which valid affect diagnosis facilitates adaptation of tutoring. Further research should investigate methodologies that incorporate several of the protocols described in this chapter simultaneously, to enable a sufficient degree of convergent validity in affect measurement [48]. To this effect, [6,46] discuss various attempts to take into account the perspective of the tutor. Similarly, [56] recruited two trained teachers to provide judgments of learners' affective

states, while [52] used pedagogical experts to recommend strategies to regulate learners' affective states.

While the results of those attempts are encouraging, the development of research protocols to collect tutors' inferences is not trivial. One difficulty lies in the assumptions made about immediacy and accuracy of human reasoning [57]. To date, most efforts in the field assume that tutors respond to observable actions of the learners, thus making actions the critical features for analysis. However, as a related analysis in [6] shows this is not always true; there are many cases where tutors accumulate evidence over a series of actions. In addition, there is sometimes anticipatory behavior that is not easily captured or directly perceivable from observable actions. Case study 2 provides a first step towards taking history of a tutorial interaction into account, but further investigation reveals that even the antecedent values of reported factors also play a role in tutors' diagnosis – even for the same sequence of events, tutors' actions (and their subsequent verbalizations and reports) are affected by values reported earlier for the same factor [46].

A possible solution that is emerging is to compare, aggregate and consolidate models developed from different sources of data (e.g. self-reports and peer or tutor reports). On the one hand, learners are a more valid source of evidence for reporting their own affective states such as their level of confidence, than the tutors. On the other hand, tutors may be better suited than the learners to judge learners' boredom or effort as well as to report on how such judgments can be used to support learning. Actor-observer biases may also play an important role in the judgments of fuzzy, vague, ill-defined constructs such as affective states. Learners might provide one set of categories by attributing their states to situational factors, while observers (peers and trained judges) might make attributions to stable dispositional factors, thereby obtaining an alternate set of categories [58].

Despite these qualifications, examples of how to derive models from different sources of data, can be found in [6] where different branches of decisions trees are manually aggregated. Similar

examples appear in [50]. If done using automatic methods, this approach has the potential to increase the precision of the models generated. The issue of automatically aggregating models automatically has been investigated in detail in the field of data mining [59,60]. In addition, the need to consider and merge different perspectives resembles the emerging requirements behind reconciling models in the field of ontologies e.g. [61,62]. Insights of how this could be achieved appear in [63]. A particularly relevant example is the work presented in [64] where a user's and an expert's conceptual model are compared. Developing formal ways to perform such measurements is necessary to enable the reduction of the bias introduced by researchers' intuitions.

V. CONCLUSIONS

As illustrated by the case studies, monitoring students' interaction parameters can provide a cost-effective, non-intrusive, efficient, and effective method to automatically detect complex phenomenon such as the affective states that accompany learning. However, several methodological issues emerged and were discussed in this chapter. One important aspect to consider was the context of data collection and the inevitable interference created by affective labels collection. Another important issue that was flagged is students' familiarity with the learning environment and their goals when using it. Finally, two methods of monitoring affective states were described and employed in the case studies: a) real-time measurements by means of emote-aloud protocols or observations and b) retrospective affect judgment by participant and/or tutors.

Although the studies used specific systems and measured certain affective states, the methodology for data collection and the machine learning methods employed are generalisable and could serve as guidance for the design of other studies. However, several important questions still remain. There is the question of how these features can be coupled with the bodily measures such as facial features, speech contours, and body language, as well as physiological measures such as galvanic skin response, heart rate, respiration rate, etc. Detection accuracies could be increased by implementing hybrid models and consolidating their outputs. There is also the question of how computer tutors might adapt their pedagogical and motivational strategies to be responsive to assessments of learners' affective

and cognitive states to heighten engagement and optimize learning gains. The accuracy as well as the type of response has to be interpreted in the context of the likely educational consequences of incorrect predictions. It is clear that although computer tutors that respond in this fashion represent a significant advance over ITSs that are mere cognitive machines; these advances await further research and technological development.

ACKNOWLEDGMENTS

D'Mello and Graesser would like to acknowledge the National Science Foundation (REC 0106965, ITR 0325428, HCC 0834847) for funding this research. Any opinions, findings and conclusions, or recommendations expressed in this chapter are those of the authors and do not necessarily reflect the views of NSF.

REFERENCES

1. Arroyo, I., Cooper, D., Burleson, W., Woolf, B., Muldner, K., and Christopherson, R., Emotion Sensors Go To School, in *Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling, Proceedings of the 14th International Conference on Artificial Intelligence in Education,*, Dimitrova, V., Mizoguchi, R., du Boulay, B., and Graesser, A. IOS Press,vol 200, pp. 17-24. 2009

2. Forbes-Riley, K., Rotaru, M., and Litman, D., The relative impact of student affect on performance models in a spoken dialogue tutoring system, *User Modeling and User-Adapted Interaction* 18 (1), pp. 11-43, 2008.

3. Conati, C. and Maclaren, H., Empirically building and evaluating a probabilistic model of user affect, *User Modeling and User-Adapted Interaction* 19 (3), pp. 267-303, 2009.

4. Robison, J., McQuiggan, S., and Lester, J., Evaluating the Consequences of Affective Feedback in Intelligent Tutoring Systems, in *International Conference on Affective Computing & Intelligent Interaction*, Amsterdam, pp. 1-6. 2009

5. D'Mello, S., Craig, S., Fike, K., Graesser, A., and Jacko, J., Responding to Learners' Cognitive-Affective States with Supportive and Shakeup Dialogues, *Human-Computer Interaction. Ambient, Ubiquitous and Intelligent Interaction*, Springer Berlin Heidelberg, 2009, pp. 595-604.

6. Porayska-Pomsta, K., Mavrikis, M., and Pain, H., Diagnosing and acting on student affect: the tutor's perspective, *User Modeling and User-Adapted Interaction* 18 (1), pp. 125-173, 2008.

7. Lepper, M. R., Woolverton, M., Mumme, D. L., Gurtner, J., Lajoie, S. P., and Derry, S. J., Motivational Techniques of Expert Human Tutors: Lessons for the Design of Computer-Based Tutors, *Computers as Cognitive Tools*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1993, pp. 75-107.

8. Goleman, D., *Emotional intelligence: Why it can Matter more than IQ* Boomsbury, 1996.

9. Conati, C., Marsella, S., and Paiva, A., Affective interactions: the computer in the affective loop, in *Proceedings of the 10th international conference on Intelligent User Interfaces* ACM Press, San Diego, California, USA, p 7. 2005
10. Conati, C., Probabilistic Assessment of User's Emotions in Educational Games, *Journal of Applied Artificial Intelligence* 16 (7-8), pp. 555-575, 2002.
11. Carver, C., Negative Affects Deriving From the Behavioral Approach System, *Emotion* 4 (1), pp. 3-22, 2004.
12. Deci, E. L., Ryan, R. M., and Aronson, J., The paradox of achievement: The harder you push, the worse it gets, *Improving academic achievement: Impact of psychological factors on education*, Academic Press, Orlando, FL, 2002, pp. 61-87.
13. Dweck, C. S. and Aronson, J., Messages that motivate: How praise molds students' beliefs, motivation, and performance (in surprising ways), in Aronson, J. (Ed.),*Improving academic achievement: Impact of psychological factors on education*, Academic Press, New York, 2002.
14. Stein, N. L., Hernandez, M. W., Trabasso, T., Lewis, M., Haviland-Jones, J. M., and Barrett, L. F., Advances in modeling emotions and thought: The importance of developmental, online, and multilevel analysis, *Handbook of emotions*, Guilford Press, New York, 2008, pp. 574-586.
15. Keller, J. M. and Reigeluth, C. M., Motivational design of instruction, *Instructional-design Theories and Models: an Overview of their Current Status*, Laurence Erlbaum Associates Hillside New Jersey, 1983, pp. 383-434.
16. Ames, Classrooms: goals, structures, and student motivation, *Journal of Educational Psychology* 84 (3), pp. 261-271, 1992.
17. Rosiek, J., Emotional Scaffolding: An Exploration of the Teacher Knowledge at the Intersection of Student Emotion and the Subject Matter, *Journal of Teacher Education* 54 (5), pp. 399-412, 2003.
18. Qu, L., Wang, N., and Johnson, L., Using Learner Focus of Attention to Detect Learner Motivation Factors, in *Proceedings of the User Modeling Conference 2005* pp. 70-73. 2005
19. Beck, J., Engagement tracing: using response times to model student disengagement, in *Proceeding of the 2005 conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology* pp. 88-95. 2005
20. Johns, J. and Woolf, P., A Dynamic Mixture Model to Detect Student Motivation and Proficiency, in *AAAI* pp. 163-168. 2006
21. Ryan Shaun Joazeiro de, B., Corbett, A., Roll, I., and Koedinger, K., Developing a generalizable detector of when students game the system, *User Model. User-Adapt. Interact.* 18 (3), pp. 287-314, 2008.
22. Walonoski, J. and Heffernan, N., Detection and Analysis of Off-Task Gaming Behavior in Intelligent Tutoring Systems, in *Proceedings of the 8th Conference on Intelligent Tutoring Systems* pp. 382-391. 2006
23. Arroyo, I. and Woolf, B., Inferring learning and attitudes with a Bayesian Network of log files data, in *Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology (Proceedings of AIED-2005 Conference, 18-22 July, 2005, Amsterdam)*, Looi, C. K., McCalla, G., Bredeweg, B., and Breuker, J. IOS Press, pp. 33-40. 2005
24. Cocea, M. and Weibelzahl, S., Eliciting Motivation Knowledge from Log Files Towards Motivation Diagnosis for Adaptive Systems, in *Proceedings of the 11th International Conference on User Modeling 2007* pp. 197-206. 2007
25. Picard and Scheirer, J., The Galvactivator: A Glove that Senses and Communicates Skin Conductivity, in *Proceedings of the 9th International Conference on HCI* pp. 1538-1542. 2001
26. Pantic, M. and Rothkrantz, L., Toward an affect-sensitive multimodal human-computer interaction,

Proceedings of the IEEE 91 (9), pp. 1370-1390, 2003.
27. Zeng, Z., Pantic, M., Roisman, G., and Huang, T., A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31 (1), pp. 39-58, 2009.
28. Kapoor, A., Picard, R. W., and Ivanov, Y., Probabilistic Combination of Multiple Modalities to Detect Interest, in *International Conference on Pattern Recognition* pp. 969-972. 2004
29. Messom, C. H., Sarrafzadeh, A., Johnson, M. J., and Chao, F., Affective state estimation from facial images using neural networks and fuzzy logic, in *Neural Networks Applications in Information Technology and Web Engineering*, Wang, D.2005
30. Litman, Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors, *Speech communication* 28 (5), pp. 559-590, 2006.
31. D'Mello, S., Graesser, A., and Picard, R. W., Toward an Affect-Sensitive AutoTutor, *Intelligent Systems, IEEE* 22 (4), pp. 53-61, 2007.
32. Ang, J., Dhillon, R., Krupski, A., Shriberg, E., and Stolcke, A., Prosody-based automatic detection of annoyance and frustration in human-computer dialog, in *Proccedings of the International Conference on Spoken Language Processing*, Denver, CO.,vol 3, pp. 2037-2039. 2002
33. Forbes-Riley, K. and Litman, D. J., Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources, in *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)* pp. 201-208. 2004
34. Liscombe, J., Riccardi, G., and Hakkani-Tür, D., Using Context to Improve Emotion Detection in Spoken Dialog Systems, in *9th European Conference on Speech Communication and Technology (EUROSPEECH'05)* Lisbon, Portugal., 2005, pp. 1845-1848.
35. Barrett, F., Are Emotions Natural Kinds?, *Perspectives on Psychological Science* 1, pp. 28-58, 2006.
36. Aviezer, H., Ran, H., Ryan, J., Grady, C., Susskind, J. M., Anderson, A. K., and Moscovitch, M., Angry, Disgusted or Afraid?, *Studies on the Malleability of Facial Expression Perception* 19 (7), pp. 724-732, 2008.
37. Russell, J. A., Core affect and the psychological construction of emotion, *Psychological review* 110 (1), pp. 145-172, 2003.
38. Stemmler, G., Heldmann, M., Pauls, C. A., and Scherer, T., Constraints for emotion specificity in fear and anger: the context counts, *Psychophysiology* 38 (2), pp. 275-291, 2001.
39. D'Mello, S. K., Craig, S. D., Sullins, C. J., and Graesser, A. C., Predicting Affective States expressed through an Emote-Aloud Procedure from AutoTutor's Mixed Initiative Dialogue, *International Journal of Artificial Intelligence in Education* 16 (1), pp. 3-28, 2006.
40. Graesser, A. C., Chipman, P., Haynes, B. C., and Olney, A., AutoTutor: an intelligent tutoring system with mixed-initiative dialogue, *IEEE Transactions on Education* 48 (4), pp. 612-618, 2005.
41. Ericsson, K. A. and Simon, H. A., *Protocol Analysis: Verbal Reports as Data* MIT Press, Cambridge, MA, 1993.
42. Trabasso, T. and Magliano, J. P., Conscious understanding during comprehension, *Discourse Processes* 21 (3), pp. 255-287, 1996.
43. Baker, R., Rodrigo, M., and Xolocotzin, U., The dynamics of affective transitions in simulation problem-solving environments, in Paiva, A. P. R. P. R. W. (Ed.),*2nd International Conference on Affective Computing and Intellegent Interaction*2007, pp. 666-677.
44. Conati, C., Chabbal, R., and Maclaren, H., A Study on Using Biometric Sensors for Monitoring User Emotions in Educational Games, in *Workshop on Assessing and Adapting to User Attitudes and Affect: Why, When and How? in conjunction with User Modeling (UM-03*, Johnstown, USA,2003

45. Mavrikis, M., Maciocia, A., and Lee, J., Towards Predictive Modelling of Student Affect from Web-Based Interactions, in *Artificial Intelligence in Education: Building Technology Rich Learning Contexts that Work (Proceedings of the 13th International Conference on Artificial Intelligence in Education, AIED2007)*, Luckin, R., Koedinger, K., and Greer, J. IOS Press,vol 158, pp. 169-176. 2007
46. Mavrikis, M., *Modelling Students' Behaviour and Affective States in ILEs through Educational Data Mining*, PhD Thesis, The University of Edinburgh, 2008.
47. Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling, Proceedings of the 14th International Conference on Artificial Intelligence in Education, AIED 2009, July 6-10, 2009, Brighton, UK, in *AIED*, Dimitrova, V., Mizoguchi, R., du Boulay, B., and Graesser, A. IOS Press,vol 200,2009
48. D'Mello, S., Craig, S., and Graesser, A., Multimethod assessment of affective experience and expression during deep learning, *International Journal of Learning Technology* 4 (3/4), pp. 165-187, 2009.
49. Baker, R. S., Corbett, A. T., Koedinger, K. R., and Wagner, A. Z., Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game The System", in *Proceedings of ACM CHI 2004: Computer-Human Interaction* pp. 383-390. 2004
50. Graesser, A. C., McDaniel, B., Chipman, P., Witherspoon, A., D'Mello, S., and Gholson, B., Detection of Emotions during learning with AutoTutor, in *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, Mahwah, NJ, pp. 285-290. 2006
51. Landauer, T. and Dumais, S., A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge, *Psychological Review* 104 (2), pp. 211-240, 1997.
52. D'Mello, S., Craig, S., Witherspoon, A., McDaniel, B., and Graesser, A., Automatic detection of learner's affect from conversational cues, *User Modeling and User-Adapted Interaction* 18 (1-2), pp. 45-80, 2008.
53. Mavrikis, M. and Maciocia, A., WALLIS: a Web-based ILE for Science and Engineering Students Studying Mathematics, *Workshop of Advanced Technologies for Mathematics Education in 11th International Conference on Artificial Intelligence in Education, Sydney Australia*, 2003.
54. Lohr, S., *Sampling: Design and Analysis* Duxbury Press, 1999.
55. Witten, I. and Frank, E., *Data Mining: Practical machine learning tools and techniques* Morgan Kaufmann, San Francisco, 2005.
56. D'Mello, S., Taylor, R., Davidson, K., and Graesser, A., Self versus teacher judgments of learner emotions during a tutoring session with AutoTutor, in *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, Woolf, B. P., Aimeur, E., Nkambou, R., and Lajoie, S., Montreal, CANADA, pp. 9-18. 2008
57. Porayska-Pomsta, K., *Influence of Situational Context on Language Production: Modelling Teachers' Corrective Responses*, PhD thesis, School of Informatics, The University of Edinburgh, 2003.
58. Jones, E. and Nisbett, R., *The Actor and the Observer: Divergent Perceptions of the Causes of Behavior* General Learning Press, New York, 1971.
59. Williams, G. J., *Inducing and Combining Decision Structures for Expert Systems*, The Australian National University, 1990.
60. Vannoorenberghe, P., On aggregating belief decision trees, *Information Fusion* 5 (3), pp. 179-188, 2004.
61. Ehrig, M. and Sure, Y., Ontology mapping - an integrated approach, in *European Semantic Web Symposium (ESWS)* pp. 76-91. 2004
62. Klein, M., Combining and relating ontologies: an analysis of problems and solutions, in *Workshop*

on Ontologies and Information Sharing, IJCAI'01, Perez, G., Gruninger, M., Stuckenschmidt, H., and Uschold, M.2001
63. Agarwal, P., Huang, Y., and Dimitrova, V., Formal Approach to Reconciliation of Individual Ontologies for Personalisation of Geospatial Semantic Web, in *Proceedings of GeoS 2005, LNCS 3799*, Rodriguez, M. A.2005, pp. 195-210.
64. Arroyo, A., Denaux, R., Dimitrova, V., and Pye, M., Interactive Ontology-Based User Knowledge Acquisition: A Case Study, in *Semantic Web: Research and Applications, Proceedings of the 3rd European Semantic Web Conference, ESWC 2006*, Sure, Y. and Dominguez, J. pp. 560-574. 2006

confidence.png