Human Expert Labeling Process: Valence-Arousal Labeling for Students' Affective States

Sinem Aslan[1], Eda Okur[1], Nese Alyuz[1], Asli Arslan Esme[1], Ryan S. Baker[2]

[1] Intel Corporation, Hillsboro OR 97124, USA
{sinem.aslan, eda.okur, nese.alyuz.civitci, asli.arslan.esme}@intel.com
[2] University of Pennsylvania, Philadelphia PA 19104, USA
rybaker@upenn.edu

Abstract. Affect has emerged as an important part of the interaction between learners and computers, with important implications for learning outcomes. As a result, it has emerged as an important area of research within learning analytics. Reliable and valid data labeling is a key tenet for training machine learning models providing such analytics. In this study, using Human Expert Labeling Process (HELP) as a baseline labeling protocol, we investigated an optimized method through several experiments for labeling student affect based on Circumplex Model of Emotion (Valence-Arousal). Using the optimized method, we then had the human experts label a larger quantity of student data so that we could test and validate this method on a relatively larger and different dataset. The results showed that using the optimized method, the experts were able to achieve an acceptable consensus in labeling outcomes as aligned with affect labeling literature.

Keywords: Affective State Labeling, Circumplex Model of Emotion, Inter-Rater Agreement, Intelligent Tutoring Systems, Affective Computing.

1 Introduction

Affect has emerged as an important part of the interaction between learners and computers, with important implications for learning and learner outcomes. As a result, it has emerged as an important area of research within learning analytics [1-3]. Reliable and valid data labeling is a key tenet for training machine learning models providing such analytics.

However, there is still considerable disagreement on key aspects of the study of affect - including even how affect itself is conceptualized. Two key paradigms have emerged for how affect is represented by researchers: (1) Affect as a set of discrete states [4-9] and (2) affect as a combination of a two-dimensional space of attributes. Several models exist that represent affect and emotion as a set of discrete states; perhaps the most widely-known such model is Ekman's set of six basic emotions [10], but other key models include the OCC model of the cognitive structure of emotions [11], and the set of affective states studied by D'Mello and Graesser and their colleagues [12]. On

the other hand, various models exist that represent affect as a two-dimensional structure [13]. The most widely-used model, however, both in education and other domains, is Russell's Circumplex Model of Emotion [14], which represents affect as a 2x2 combination of Valence (Negative to Positive) and Arousal (Calm to Excited).

Human Expert Labeling Process (HELP) is a labeling protocol [15], which was originally developed to enable affect labelers (i.e., human experts) with backgrounds in Psychology or Educational Psychology label students' discrete affective states (i.e., Satisfied, Confused, and Bored) occurring in a 1:1 digital learning scenario. In this study, using HELP as a baseline labeling protocol, we investigated an optimized method for labeling student affect based on Circumplex Model of Emotion (Valence-Arousal). Therefore, there is one major research question that this study aims to address: What is an optimized method for labeling students' affect in terms of valence and arousal? Identification of such a method will be critical for obtaining ground truths necessary for generation of analytics based on machine learning techniques.

2 Data Collection

The student data used in this study is a part of a larger dataset previously collected through authentic classroom pilots of an afterschool Math course in an urban high school in Turkey [16]. During the pilots, the students used an online learning platform for watching instructional videos and solving assessment questions. Our data collection application running in the background collected two streams of videos from the students: (1) Student appearance videos from the camera, to enable monitoring of observable cues available in the individual's face or upper body; and (2) student desktop videos, to enable observation of contextual information.

3 Labeling Tool, Labelers, and Training

We developed a labeling tool customized to various labeling experiments (see Fig. 1).

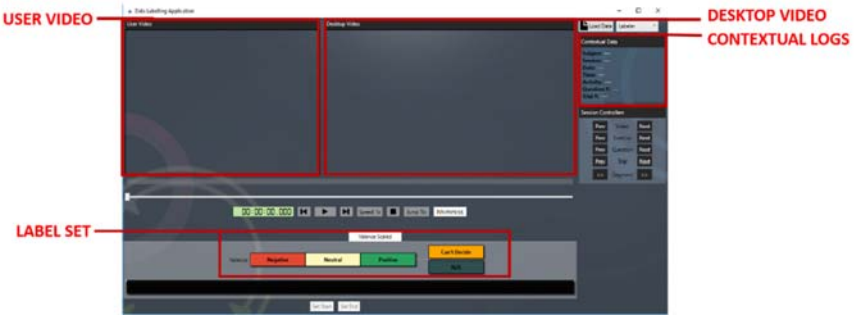

Fig. 1. Customized labeling tool (sample view), for labeling Valence.

We recruited and trained six human experts with backgrounds in Psychology/Educational Psychology. The training process took around eight hours which included instruction and demonstration, practice with feedback, as well as reviewing each other's labels and discussing differences in labeling outcomes. For all labeling tasks, based on observed state changes, the experts provided their Valence-Arousal labels using all available cues (e.g., student video/audio, desktop recording with mouse cursor locations, as well as any relevant contextual information from the device and content platform).

4 Experimental Conditions

To find an optimized method for Valence-Arousal labeling on the student data, we experimented with two variables: (1) Selection of Labels - which labels to use; and (2) Labeling Method - how to label.

For Selection of Labels, we had two conditions: Binary Labeling vs. Scaled Labeling. Binary Labeling had two levels of states: Positive vs. Negative for Valence, and Low vs. High for Arousal. Scaled Labeling had a scale of three levels: Negative, Neutral, and Positive for Valence; and Low, Medium, and High for Arousal.

For Labeling Method, we had three conditions: (1) Separate Labeling, (2) Combined Labeling, and (3) Separate Labeling with Displayed Labels. In the Separate Labeling condition, the human experts were asked to label either Valence or Arousal - one at a time. In Combined Labeling condition, they were asked to label Valence and Arousal simultaneously. In Separate Labeling with Displayed Labels condition, the experts labeled one construct first, and then labeled the other construct with the first construct's labels displayed - i.e., the experts were asked to label either Valence with their previous label of Arousal for the same data displayed or label Arousal with their previous labels of Valence for the same data displayed.

For Selection of Labels, we assigned the human experts to the conditions (Binary vs. Scaled) at the beginning of the study so that we could train them based on the specific labels for their assigned condition: We randomly assigned three experts to the Binary Labeling condition, and the other three to the Scaled Labeling condition.

For all three Labeling Method conditions (i.e., Separate Labeling, Combined Labeling, and Separate Labeling with Displayed Labels), the human experts in both Binary and Scaled Labeling individually labeled the same student data (around seven hours collected from five students in two sessions - each session was 40 minutes). In summary, all six experts followed the procedures outlined below (each expert labeled the same student data five times so that we could have comparative results):

1. Valence labeling only (Separate Labeling).
2. Arousal labeling only (Separate Labeling).
3. Valence and Arousal labeling together (Combined Labeling).
4. Arousal labeling with Valence labels displayed (Separate Labeling with Displayed Labels).
5. Valence labeling with Arousal labels displayed (Separate Labeling with Displayed Labels).

Note that we randomized the order of the first three procedures to minimize the effect of time and familiarity of the student data being labeled. After the experts completed (1-3), they conducted (4) first, and then (5). Note that (4) and (5) were conducted after (1-3), since we needed either Valence or Arousal labels gathered from the individual experts so that we could display them during labeling.

5 Valence-Arousal Labels

In this research, Valence is defined as the direction of a student's affect and Arousal as the level of activation in physical response of the student during the learning process. For Valence, we had three possible labels:

1. Negative: The student seems to experience negative affect (e.g., getting frustrated, stressed, agitated, bored, etc.). Any negative affect is placed within this category.
2. Neutral: The student's affect seems to be neutral. One cannot observe any clear direction towards negative or positive affect (e.g., calm).
3. Positive: The student seems to experience positive affect (e.g., feeling satisfied, excited, etc.). Any positive affect is placed within this category.

For Arousal, we had three possible labels:

1. Low: The student does not seem to be emotionally activated, dynamic, reactive, or expressive of his/her affect.
2. Medium: The student seems to be emotionally somewhat dynamic, reactive, and expressive of his/her affect.
3. High: The student seems to be emotionally very dynamic, reactive, and expressive of his/her affect.

In Table 1, we summarized the final list of Valence-Arousal labels as customized for the Binary and Scaled Labeling conditions. In addition to Valence and Arousal labels, we also had control labels that apply to both of these conditions: Can't Decide (if the human expert cannot decide on a final label) and N/A (if data cannot be labeled - e.g., there is no one in front of the camera).

Table 1. Binary and Scaled Valence-Arousal Labels

	Binary Labels	Scaled Labels
Valence	Negative vs. Positive	Negative–Neutral–Positive
Arousal	Low vs. High	Low–Medium–High

6 Analysis of the Labeled Data

Upon completion of labeling, we preprocessed the labeled data prior to analysis: We first aligned label-sets of all experts to each other. Then, we applied windowing over

each expert's labeling outputs to obtain the corresponding instance-wise labels. For this, we utilized a sliding window of 8 seconds with an overlap of 4 seconds. Hence, after preprocessing, we obtained instance-wise label sets that were timely synchronized with each other.

To compare labeling results for different experimental conditions, we calculated inter-rater agreement among multiple human experts. For inter-rater agreement, we used consensus measures which are designed to estimate the degree of agreement among multiple experts [18]. In this study, we used Krippendorff's alpha [19], as it is robust against incomplete data and is suitable for multiple raters. Despite of disagreements for acceptable value in the related literature, a value above 0.4 is often considered moderate agreement for affect labeling [20].

To investigate the differences among different experimental conditions (i.e., Set of Labels and Labeling Method), inter-rater agreement measures were calculated for the given Valence-Arousal labels using the indicated labeling method:

- Valence Labels with Separate Labeling
- Arousal Labels with Separate Labeling
- Valence Labels with Combined Labeling
- Arousal Labels with Combined Labeling
- Arousal Labels with Arousal Labeling with Valence Labels Displayed
- Valence Labels with Valence Labeling with Arousal Labels Displayed

All these analyses were conducted for the Binary and Scaled label sets separately. Furthermore, we conducted additional analysis, to provide a comparison between Binary and Scaled results: We post-processed Scaled label sets, converting Neutral/Medium labels to either extreme (checking both possibilities), to obtain pseudo-Binary labels. See below for how we converted these labels and their acronyms as used in the Results section:

- For Valence: **NN**: Negative and Neutral merged. | **NP**: Neutral and Positive merged.
- For Arousal: **LM**: Low and Medium merged. | **MH**: Medium and High merged.

7 Results

The inter-rater agreement results for each experimental conditions are summarized in Table 2. These results show that for Valence, the highest consensus among the human experts was achieved in the Separate Labeling with Binary Labels condition (0.495). However for Arousal, the best consensus was obtained when the experts used Scaled Labels (in the Arousal Labeling with Valence Displayed condition), which was followed by converting those Scaled Labels into LM Binary Labels (Low and Medium merged), obtaining an alpha of 0.602.

The findings in Table 2 also show that Valence labeling resulted in higher consensus among the experts than Arousal labeling (before any conversions into pseudo-Binary labels), regardless of whether the Binary or Scaled label set was used, for both Separate and Combined labeling. However, an exception to this finding was labeling Arousal after having previously labeled Valence, with the Valence labels displayed.

When comparing consensus among the human experts, we also found that consensus was always higher for the Binary Labeling conditions than for the Scaled conditions. This suggests that Binary Labeling was easier for the human experts. Additionally, when Scaled label sets are converted to pseudo-Binary labels, LM is always better than MH for Arousal agreements in all cases. This suggests that the experts found it more difficult to distinguish Low vs. Medium Arousal, than Medium vs. High Arousal. Similarly, when Scaled label sets are converted to pseudo-Binary labels, NP is always better than NN for Valence agreements in all cases. This implies that the experts found it more difficult to distinguish Neutral vs. Positive Valence, than Negative vs. Neutral Valence.

Table 2. Consensus (Krippendorff's alpha) among the Human Experts

		Binary Labels	**Scaled Labels**	**LM/ NN***	**MH/ NP***
Separate Labeling	Arousal	0.382	0.189	0.405	0.251
	Valence	0.495	0.225	0.266	0.393
Combined Labeling	Arousal	0.235	0.220	0.558	0.221
	Valence	0.355	0.237	0.333	0.388
Separate Labeling with Displayed Labels	Arousal with Valence Displayed	0.495	0.378	0.602	0.407
	Valence with Arousal Displayed	0.467	0.200	0.333	0.355

* Merging rules: LM: Low-Medium, NN: Negative-Neutral, MH: Medium-High, NP: Neutral-Positive

In addition to these quantitative results, we asked the human experts about their preferences for how to label, and which methods were easier to use, at the end of the experiments. The feedback we got from the six experts can be summarized as follows:

- The majority of the experts in the Binary labeling condition found Valence easier to label than Arousal, matching our quantitative findings.
- All experts in the Scaled labeling condition found Arousal easier than Valence to label. (Note, however, that inter-rater agreement was actually lower for Arousal in several cases, within this condition).
- 5 of the six experts indicated that they preferred to first label Valence, and then Arousal, in line with the quantitative findings.

Leveraging the quantitative results and considering the feedback from the human experts, it appears that the most optimized method for Valence-Arousal labeling on the student data, at least as far as our study is concerned, would be as follows:

1. Obtain binary Valence labels (Positive vs. Negative);
2. Displaying the binary Valence labels, obtain the scaled Arousal labels (Low-Medium-High);
3. Merge the Low and Medium Arousal scales to obtain final binary Arousal labels (Low vs. High).

Once this optimized method was identified, the next step was to test and validate this method using a relatively larger dataset from more students. Towards this end, we had the experts label around 104 hours of student data in total (from 17 students in 13 sessions) using the optimized method. The results obtained with the optimized labeling

method are summarized in Table 3, both for the complete dataset and the subset of data utilized in the previous experiments (i.e., experimental data). As the results in Table 3 indicate, consensus among the experts is even higher for the complete dataset (Valence: 0.549; Arousal: 0.610) than for the subset previously studied.

Table 3. Consensus Measures with the Optimized Method (Experimental vs. Complete Data)

	Dataset Details		Consensus Measures	
Name	Student Count	Total Number of Hours	Valence	Arousal
Experimental	5	7	0.495	0.602
Complete	17	104	0.549	0.610

8 Conclusions

To enable the human experts conduct Valence-Arousal labeling on the student data, we used HELP as a baseline labeling protocol. However, as the protocol was originally developed for labeling discrete affective states only, we needed to identify an optimized method for Valence-Arousal labeling on the student data through several experiments. Empirically, the optimized labeling methodology was found to be consisting of the following steps: (1) Obtaining binary Valence labels, (2) obtaining scaled Arousal labels when binary Valence labels are displayed, and (3) merging Low and Medium Arousal scales to obtain final binary Arousal labels. Additionally, the results of our experiments suggest that, before pseudo-Binary conversions, Valence labeling was easier than Arousal labeling for the human. The only exception to this finding was labeling Arousal with the Valence labels displayed, and this may be because labeling Arousal was generally more difficult, and labeling Valence first could have helped the experts isolate their thinking about Arousal, not taking Valence into account.

Using the optimized method, we then had the human experts label a larger quantity of student data so that we could test and validate this method on a relatively larger and different student dataset. The results showed that using the optimized method, the experts were able to achieve an acceptable consensus in labeling outcomes as aligned with relevant affect labeling literature [15, 17]. The researchers and practitioners can leverage the results of this study to design and implement similar data labeling tasks with a consideration of some limitations of the study (e.g., limited number of students, education-context dependency).

References

1. Sabourin, J., Mott, B., Lester, J. C.: Modeling learner affect with theoretically grounded dynamic Bayesian networks. In: Proceedings of International Conference on Affective Computing and Intelligent Interaction, 286-295. Springer, Berlin, Heidelberg (2011).
2. Jaques, N., Conati, C., Harley, J. M., Azevedo, R.: Predicting affect from gaze data during interaction with an intelligent tutoring system. In: Proceedings of the International Conference on Intelligent Tutoring Systems, 29-38. Springer, Cham (2014).

3. Pardos, Z.A., Baker, R.S., San Pedro, M.O.C.Z., Gowda, S.M., Gowda, S.M.: Affective states and state tests: Investigating how affect and engagement during the school year predict end of year learning outcomes. Journal of Learning Analytics, 1(1), 107-128 (2014).
4. Kapoor A., Picard, R. W.: Multimodal affect recognition in learning environments. In: Proceedings of the International Conference on Multimedia, 677-682. ACM (2005).
5. Kapoor, A., Burleson, W., Picard, R. W.: Automatic prediction of frustration. International Journal of Human-Computer Studies, 65(8), 724-736 (2007).
6. Hoque, M. E., McDuff, D. J., Picard, R. W.: Exploring temporal patterns in classifying frustrated and delighted smiles. Transactions on Affective Computing, 65(8), 323-334 (2012).
7. Grafsgaard, J. F., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., Lester, J. C.: Automatically recognizing facial indicators of frustration: a learning-centric analysis. In: Proceedings of the International Conference on Affective Computing and Intelligent Interaction, 159-165, IEEE (2013).
8. Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V.,Ventura, M., Zhao, W.: Automatic detection of learning centered afective states in the wild. In: Proceedings of the International Conference on Intelligent User Interfaces, 379-388. ACM (2015).
9. Arroyo, I., Cooper, D. G., Burleson, W., Woolf, B. P., Muldner, K., Christopherson, R.: Emotion sensors go to school. In: Proceedings of the International Conference on Artificial Intelligence in Education, 200, 17-24 (2009).
10. Ekman, P.: An argument for basic emotions. Cognition & emotion, 6(3-4), 169-200 (1992).
11. Ortony, A., Clore, G.L., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press, Cambridge, UK (1988).
12. D'Mello, S., Picard, R.W., Graesser, A.: Toward an affect-sensitive AutoTutor. Intelligent Systems, 22(4). IEEE (2007).
13. Barrett, L. F., Russell, J. A.: The structure of current affect: Controversies and emerging consensus. Current Directions in Psychological Science, 8(1), 10-14 (1999).
14. Russell, J. A.: A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161 (1980).
15. Aslan, S., Mete, S. E., Okur, E., Oktay, E., Alyuz, N., Genc, U., Stanhill, D., Arslan Esme, A.: Human expert labeling process (HELP): towards a reliable higher-order user state labeling process and tool to assess student engagement. Educational Technology, 57(1), 53-59 (2017).
16. Okur, E., Alyuz, N., Aslan, S., Genc, U., Tanriover, C., Arslan Esme, A.: Behavioral engagement detection of students in the wild. In: Proceedings of the International Conference on Artificial Intelligence in Education, 250-261. Springer, Cham (2017).
17. Ocumpaugh, J., Baker, R., Rodrigo, M. M. T.: Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) 2.0 technical and training manual. New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences (2015).
18. Stemler, S. E.: A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research and Evaluation, 9(4), 1-19 (2004).
19. Krippendorff, K.: Computing Krippendorff's alpha-reliability. Departmental Papers (ASC), 43. Retrieved from http://repository.upenn.edu/asc_papers/43 (2011).
20. Siegert, L., Böck, R., Wendemuth, A.: Inter-rater reliability for emotion annotation in human–computer interaction: comparison and methodological improvements. Journal of Multimodal User Interfaces, 8(1), 17-28 (2014).