Towards Human Affect Modeling: A Comparative Analysis of Discrete Affect and Valence-Arousal Labeling

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. There is still considerable disagreement on key aspects of affective computing - including even how affect itself is conceptualized. Using a multimodal student dataset collected while students were watching instructional videos and answering questions on a learning platform, we investigated the two key paradigms of how affect is represented through a comparative approach: (1) Affect as a set of discrete states and (2) Affect as a combination of a two-dimensional space of attributes. We specifically examined a set of discrete learning-related affects (Satisfied, Confused, and Bored) that are hypothesized to map to specific locations within the Valence-Arousal dimensions of Circumplex Model of Emotion. For each of the key paradigms, we had five human experts label student affect on the dataset. We investigated two major research questions using their labels: (1) Whether the hypothesized mappings between discrete affects and Valence-Arousal are valid and (2) whether affect labeling is more reliable with discrete affect or Valence-Arousal. Contrary to the expected, the results show that discrete labels did not directly map to Valence-Arousal quadrants in Circumplex Model of Emotion. This indicates that the experts perceived and labeled these two relatively differently. On the other side, the inter-rater agreement results show that the experts moderately agreed with each other within both paradigms. These results imply that researchers and practitioners should consider how affect information would operationally be used in an intelligent system when choosing from the two key paradigms of affect.

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

1 Introduction

Affect has become an important area of research within learning [1-3]. Data labeling is a preliminary step towards training machine learning models to provide affect-related analytics to teachers and learners. However, there is a lack of agreement in the related literature even for how affect is itself conceptualized. There are two major paradigms for affect representation: (1) Affect as a set of discrete states [4-9] and (2) Affect as a combination of a two-dimensional space of attributes [11].

There are several benefits to viewing student affect as a set of discrete states. One such benefit is easier understanding of students' actual states and driving customized interventions accordingly. However, labeling discrete affective states presents a challenge to observers in distinguishing between closely-related affective states. For instance, confusion and frustration are often treated as separate affective states (e.g., [8]), but Liu and colleagues [10] argue that they may simply represent different ranges of a continuum. Researchers using discrete sets of affective states often also struggle with how to distinguish neutral affect from mild affect and how to handle uncommon affect outside the core affect labeling scheme. These challenges can represent major risks to the quality of affect labeling in ways that are not easily seen in overall inter-rater agreement values that cut across large numbers of constructs. These issues may particularly emerge in situations where affect labelers have limited training or are asked to label data where video is sometimes ambiguous, due to factors such as facial occlusion, adverse pose variations, gum chewing, or many other factors.

In this paper, we study this issue in a focused fashion by examining a set of discrete affective states that can be reasonably expected to correlate to specific locations within the Circumplex Model of Emotion [11]. Specifically, we study (see Fig. 1): Satisfied, which can be hypothesized to map to Positive Valence (regardless of Arousal); Bored, which can be hypothesized to map to Negative Valence and Low Arousal; and Confused, which can be hypothesized to map to Negative Valence and High Arousal. Using the student dataset in [12] and Human-Expert Labeling Process (HELP) [13] as a baseline labeling protocol, we test these hypotheses (i.e., whether these mappings between discrete affective states and Valence-Arousal are valid) and if affect labeling is more reliable with discrete affective states or Valence-Arousal.

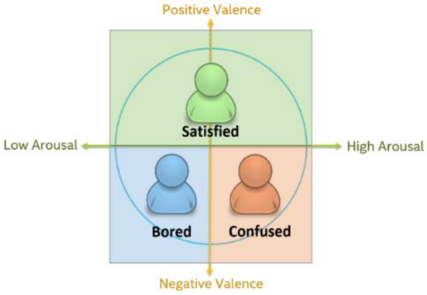

Fig. 1. Mapping of categorical emotions to the Circumplex Model of Emotion.

2 Data Collection

In this study, we used student data which was a subset of a larger dataset previously collected through authentic classroom pilots [12]. These pilots took place in an after-school Math course in an urban high school in Turkey. In these pilots, the students used an online educational platform to watch instructional videos and solve relevant questions. Meanwhile, our data collection application was running in the background to collect two video streams: (1) Student appearance videos from the camera (to monitor observable cues available in the student's face or upper body); and (2) student desktop videos (to monitor contextual information).

3 Labeling Tool, Human-Experts, and Training

A labeling tool was developed and customized for use in multiple labeling experiments. In Fig. 2, a sample view for labeling Valence is shown.

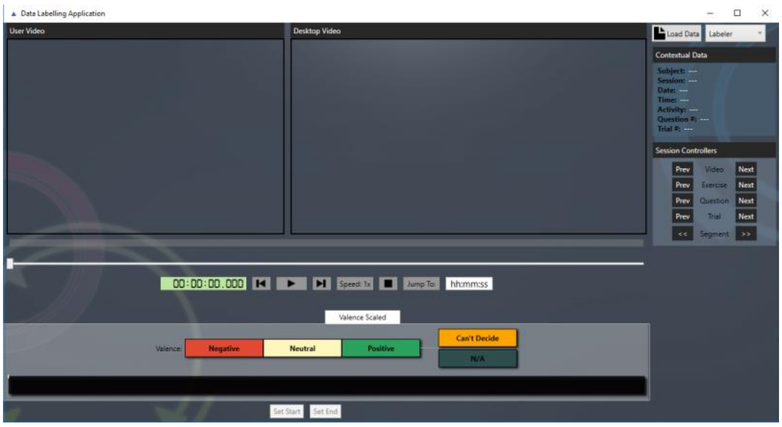

Fig. 2. Customized labeling tool (sample view), for labeling Valence.

Using HELP [13] as a baseline labeling protocol, five human experts with backgrounds in Psychology/Educational Psychology were recruited and trained (See Table 1 and 2 for operational definitions of labels). Based on observed state changes, the experts provided their Valence-Arousal or discrete affect labels using all available cues (e.g., student video/audio, desktop recording with mouse cursor locations, and any relevant contextual information from the device and content platform).

Table 1. Operational Definitions of Discrete Affect Labels

	Operational Definitions
Satisfied	If a student is not having any emotional challenges during a learning task. This can include all positive affective states from being neutral to being excited during the learning task; neutral is included here along with positive.
Confused	If the student is getting confused during the learning task – in some cases this state might include some other negative affects such as frustration (argued by [10] to represent an increased level of confusion).
Bored	If the student is feeling bored during the learning task.

Table 2. Operational Definitions of Valence-Arousal Labels

	Operational Definitions
Valence	The direction of a student's affect during the learning process with two levels: ***Positive:*** The student seems to experience neutral or positive affect (e.g., s/he is feeling calm, satisfied, excited, etc.). Any neutral or positive affect is placed within this category. ***Negative:*** The student seems to experience negative affect (e.g., s/he is getting frustrated, stressed, agitated, bored, etc.). Any negative affect is placed within this category.
Arousal	Level of activation in physical response of the student during the learning process with three levels: ***Low:*** The student does not seem to be emotionally activated, dynamic, reactive, or expressive of his/her affect. ***Medium:*** The student seems to be emotionally somewhat dynamic, reactive, and expressive of his/her affect. ***High:*** The student seems to be emotionally very dynamic, reactive, and expressive of his/her affect.

In total, the human experts labeled seven hours of student data for Valence-Arousal labels first. One week later, we asked them to label the same data for discrete affect labels. Note that although the experts labeled Arousal using three different levels, we combined Low and Medium labels into a Low class for analysis of the labeled data based on the experimental results outlined in [14].

4 Comparing Discrete Affect Labels to Valence-Arousal Labels

4.1 Pre-processing of Label Data

To analyze labeling output data, both for discrete affect and Valence-Arousal labeling outputs, two pre-processing steps were taken: First, we applied windowing on the labeling output data to obtain aligned instance-wise labels of each individual expert. Second, to facilitate analysis, we derived a consensus label from all the expert labels for each instance, using majority voting in each case.

4.2 Metrics for Analysis

The derived consensus labels were then correlated to each other to measure the degree to which each discrete affective state mapped to each Valence-Arousal quadrant. Note that we already presented the hypotheses for how discrete affective states would map to Valence-Arousal in the Introduction section (Fig. 1). We calculated the degree of mapping using Precision, Recall, and F1-measures. For these calculations, the labeled set (e.g., discrete affective states) act as the true labels; whereas the mapped set (e.g., Valence-Arousal mapped to discrete affective states as hypothesized) serve as the predictions. Precision is calculated as the fraction of true predictions (i.e., true positives) to the number of all predictions (i.e., sum of true positives and false positives); whereas recall is calculated as the ratio of true predictions to all true labels (i.e., sum of true positives and false negatives). The F1 measure is calculated as the harmonic mean of precision and recall values, taking into account the trade-off between those two measures. In addition, we also checked inter-rater agreement measures for different labeling tasks (i.e., Discrete Affects, Arousal, Valence) to assess reliability of the obtained label data. As proposed in HELP [13], we utilized Krippendorff's alpha metric to compute inter-rater agreement among experts.

4.3 Methods for Analysis

To investigate whether the discrete affective states (i.e., Satisfied, Bored, and Confused) actually map to the hypothesized Valence-Arousal quadrants, the degree of mappings was computed using the final labels for the following mapping/comparison sets:

- Valence vs. Discrete Affect-to-Valence: We compared Valence labels to discrete affect labels, where affect labels were mapped to Valence labels using: Satisfied to Positive Valence, and Bored/Confused to Negative Valence.
- Arousal vs. Discrete Affect-to-Arousal: We compared Arousal labels to discrete affect labels, where affect labels were mapped to Arousal labels using: Bored to Low Arousal, and Confused to High Arousal. Note that Satisfied samples were disregarded in this case since we hypothesized that they could map to both Low and High Arousal on the Circumplex Model of Emotion (See Fig. 1).
- Discrete Affect vs. Valence/Arousal-to-Discrete Affect: We compared discrete affect labels to Valence-Arousal labels, where Valence-Arousal label pairs were mapped to discrete affect labels using: Low/High Arousal & Positive Valence to Satisfied, Low Arousal & Negative Valence to Bored, and High Arousal & Negative Valence to Confused.

5 Results

5.1 Mapping between Discrete Affect and Valence-Arousal Labels

The Precision, Recall, and F1-measures calculated for each mapping sets are summarized in Table 3. As these results indicate, relatively higher F1 measures (consistent for both state-specific and overall results) could be achieved when discrete affect labels were mapped to Positive/Negative Valence (i.e., Valence vs. Discrete Affect-to-Valence). However, the F1 values were lower when discrete affect labels were mapped to High/Low Arousal (i.e., Arousal vs. Discrete Affect-to-Arousal). Although the overall F1 measures seemed reasonable when Valence-Arousal labels were mapped to discrete affects (i.e., Discrete Affect vs. Valence/Arousal-to-Discrete Affect), the state-specific measures highlighted the inconsistency. The reason behind that could be the fact that the distribution of High Arousal samples was lower than ~1.2% in the data, and the samples that were labeled as Confused were therefore drawn mostly from the Low-Arousal samples. This issue was mostly visible when we investigate the Valence-Arousal vs. Discrete Affect mapping Recall and F1 results. Note that although we disregarded Satisfied samples in Arousal vs. Discrete Affect-to-Arousal case with the hypothesis that they could map to both Low and High Arousal, we also checked and observed that among all the Satisfied instances, 99.8% are mapping to Low Arousal and only 0.2% are mapping to High Arousal. Note that this issue is common in all three discrete affective states: Satisfied (0.2% High Arousal), Bored (2.2% High Arousal), and Confused (3.3% High Arousal).

Table 3. Precision / Recall / F1 Measures for the Mappings between Discrete Affect Labels and Valence-Arousal Labels

Mapping/Comparison Set	Precision	Recall	F1
Valence vs. Discrete Affect-to-Valence			
Positive (Satisfied)	0.99	0.71	0.82
Negative (Bored/Confused)	0.41	0.96	0.57
Overall	0.75	0.75	0.75
Arousal vs. Discrete Affect-to-Arousal			
Low (Bored)	0.98	0.49	0.65
High (Confused)	0.03	0.64	0.07
Overall	0.49	0.49	0.49
Discrete Affect vs. Valence/Arousal-to-Discrete Affect			
Satisfied (Low/High & Positive)	0.71	0.99	0.83
Bored (Low & Negative)	0.75	0.51	0.61
Confused (High & Negative)	0.73	0.02	0.04
Overall	0.72	0.72	0.72

5.2 Inter-rater Agreement for Discrete Affects and Valence-Arousal Labeling

The inter-rater agreement results for discrete affect labeling compared to the Valence-Arousal labeling are given in Table 4. The average of all confusion matrices computed for discrete affect labels provided by all pairwise experts (i.e., any two expert pairs among the five experts) is given in Table 5. As these results indicate, the inter-rater agreement was lower for discrete affect labeling, where the pairwise confusion results showed that the experts had difficulty differentiating between Satisfied and any one of the other two states (Bored or Confused).

Table 4. Consensus Measures for Discrete Affects vs. Valence-Arousal

Dataset Details		Consensus Measures		
Student Count	Total Number of Hours	Valence	Arousal	Discrete Affects
5	7	0.495	0.602	0.437

Table 5. Average of Pair-wise Confusion Matrices for Discrete Affects

	Satisfied	Bored	Confused
Satisfied	1016.7	171.5	221.6
Bored	231.4	365.9	44.2
Confused	328.6	36.8	319.1

6 Conclusion

In this paper, through a comparative approach, we investigated the two key paradigms of how affect is represented: (1) Affect as a set of discrete states and (2) affect as a combination of a two-dimensional space of attributes. We specifically examined a set of discrete affective states (Satisfied, Confused, and Bored) that can be reasonably expected to map to specific locations within the Valence-Arousal dimensions of the Circumplex Model of Emotion [11]. We tested two major hypotheses: (1) Whether these mappings between discrete affects and Valence-Arousal are valid and (2) whether affect labeling is more reliable with discrete affects or Valence-Arousal. To investigate these hypotheses, we used HELP [13] as a baseline labeling protocol. Using HELP, five human experts labeled seven hours of student data for Valence-Arousal and discrete affect labels.

The relatively low F1 measures (See Table 3) indicate that the discrete affect labels (i.e., Satisfied, Bored, and Confused) do not directly map to Valence-Arousal quadrants in the Circumplex Model of Emotion [11]. This shows that the human experts perceived and labeled these two relatively differently although we reasonably expected the discrete affects to map seamlessly on the model. On the other side, the inter-rater agreement results (See Table 4) show that the experts moderately agree with each other in both discrete affect labeling and Valence-Arousal labeling.

There are two important implications of these major results to researchers in learning analytics field. First, how affect is conceptualized in one paradigm could not be seamlessly transferable to another paradigm (i.e., discrete affective states do not directly map to Valence-Arousal quadrants). Therefore, researchers need to decide on affect labels of interest at the beginning of research considering this limitation. Second, both discrete affect labeling and Valence-Arousal labeling resulted in moderate consensus among the experts. Therefore, researchers should consider how affect information would ultimately be used in a learning system (e.g., affect-aware interventions, feedback to content, etc.) when choosing from Valence-Arousal or discrete affect labeling to generate ground-truth labels for model development.

References

1. Sabourin J, Mott B, Lester JC (2011, October) Modeling learner affect with theoretically grounded dynamic Bayesian networks. In: International Conference on Affective Computing and Intelligent Interaction, Springer, Berlin, Heidelberg, pp 286-295.
2. Jaques N, Conati C, Harley JM, Azevedo R (2014, June) Predicting affect from gaze data during interaction with an intelligent tutoring system. In: International Conference on Intelligent Tutoring Systems, Springer, Cham, pp 29-38.
3. Pardos ZA, Baker RS, San Pedro MOCZ, Gowda SM, Gowda SM (2014) 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.
4. Kapoor A, Picard RW (2005) Multimodal affect recognition in learning environments. In: Int. Conf. on Multimedia.
5. Kapoor A, Burleson W, Picard RW (2007) Automatic prediction of frustration. Int. Journal of Human-Computer Studies, vol. 65, no. 8, pp 724-736.
6. Hoque ME, McDuff DJ, Picard RW (2012) Exploring temporal patterns in classifying frustrated and delighted smiles. Transactions on Affective Computing, vol. 65, no. 8, pp. 323-334.
7. Grafsgaard JF, Wiggins JB, Boyer KE, Wiebe EN, Lester JC (2013) Automatically recognizing facial indicators of frustration: a learning-centric analysis. In: Affective Computing and Intelligent Interaction.
8. Bosch N, D'Mello S, Baker R, Ocumpaugh J, Shute V, Ventura M, Zhao W (2015) Automatic detection of learning centered afective states in the wild. In: Int. Conf. on Intelligent User Interfaces.
9. Arroyo I, Cooper DG, Burleson W, Woolf BP, Muldner K, Christopherson R (2009) Emotion sensors go to school, Artificial Intelligence in Education (AIED), vol. 200, pp 17-24.
10. Liu Z, Pataranutaporn V, Ocumpaugh J, Baker RS (2013) Sequences of Frustration and Confusion, and Learning. Proceedings of the 6th International Conference on Educational Data Mining, 114-120.
11. Russell JA (1980) A circumplex model of affect. In: Journal of Personality and Social Psychology, 39, 6: 1161.
12. Okur E, Alyuz N, Aslan S, Genc U, Tanriover C, Arslan Esme A (2017, June) Behavioral engagement detection of students in the wild. In: Proceedings of the 18th International Conference on Artificial Intelligence in Education, Springer, Cham, pp 250-261.
13. Aslan S, Mete SE, Okur E, Oktay E, Alyuz N, Genc U, Stanhill D, Arslan Esme A (2017) 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.
14. Aslan S, Okur E, Alyuz N, Arslan Esme A, Baker RS (2018) Human Expert Labeling Process: Valence-Arousal labeling for students' affective states. In: Proceedings of the 8th International Conference in Methodologies and Intelligent Systems for Technology Enhanced Learning, Springer, Cham.