Don't Classify Ratings of Affect; Rank them!

Héctor P. Martínez, Georgios N. Yannakakis, *Member, IEEE,* and John Hallam

Abstract—How should affect be appropriately annotated and how should machine learning best be employed to map manifestations of affect to affect annotations? What is the use of ratings of affect for the study of affective computing and how should we treat them? These are the key questions this paper attempts to address by investigating the impact of dissimilar representations of annotated affect on the efficacy of affect modelling. In particular, we compare several different binary-class and pairwise preference representations for automatically learning from ratings of affect. The representations are compared and tested on three datasets: one synthetic dataset (testing "*in vitro*") and two affective datasets (testing "*in vivo*"). The synthetic dataset couples a number of attributes with generated rating values. The two affective datasets contain physiological and contextual user attributes, and speech attributes, respectively; these attributes are coupled with ratings of various affective and cognitive states. The main results of the paper suggest that ratings (when used) should be naturally transformed to ordinal (ranked) representations for obtaining more reliable and generalisable models of affect. The findings of this paper have a direct impact on affect annotation and modelling research but, most importantly, challenge the traditional state-of-practice in affective computing and psychometrics at large.

Index Terms—affect annotation, affect modelling, ratings, ranks, preference learning, classification, computer games, Sensitive Artificial Listener (SAL) corpus

1 Introduction

Within the very core of affect modelling research one finds the investigation of machine learning methods [1] for drawing the mapping between annotations of affective experiences and measurable variables or manifestations related to them. Ratings are one of the most popular affect annotation tools varying from simple Likert scales [2] to self-assessment manikins [3] and the rating scales of the discrete states in the Geneva emotion wheel [4]. While ratings provide solely ordinal information [5] about affective states, machine learning methods tailored for ordinal data, namely *preference learning* (PL) [6], are only rarely used in affective computing studies. Instead, classification algorithms (CL) are applied after rating annotations are transformed into a nominal representation (classes) — e.g. valence ratings are transformed into three classes: *low valence*, *medium valence* and *high valence* [7], [8], [9]. We argue that such data preprocessing practices can be detrimental to psychometrics and affective computing research efforts as they point to biased representations of affect annotation and, in turn, yield unreliable models of affect.

Alternatively, some studies use regression methods to learn the exact values of numerical ratings, specially when continuous annotations of affect are available [10], [11]. However, a number of psychometric studies shows that human ratings of emotion do not follow an absolute and consistent scale [12], [13], [14] and, thus, learning to predict ratings directly yields models of doubtful quality and use.

In this paper we show empirically that treating ratings as ordinal, instead of nominal, values generates less biased datasets and, in turn, more reliable models of affect. The paper highlights the pitfalls of classifying rating annotations compared to more adequate machine learning practices which are based on preference learning methods. Transforming ratings into classes involves a certain degree of information loss but also induces experimental biases generated via the selection of which ratings are assigned to each class. Occasionally, the transformation into classes is not even dependent on the nature of the affect modelled but on aspects linked to data analysis — e.g. splitting data accordingly to create a balanced number of samples per class (e.g. see [15]). While earlier studies have shown the benefits of rank-based (ordinal) affect annotations when compared to rating reports (e.g. see [16], [17]) this paper focuses on the impact the representation of rating annotations has on the efficiency of the affect models rather than on the annotation per se. For that purpose, we report a critical review of current affective computing practices and we provide an empirical comparison of classification and preference learning methods on rating annotations of dissimilar affective and cognitive states across different datasets.

Note that a comparison between preference learning and regression methods is not included in this paper, motivated by psychological studies suggesting that regression mistreats ratings of affect [12], [13],

- *H. P. Martínez is with the Institute of Digital Games, University of Malta, Msida, Malta. Email: hector.p.martinez@um.edu.mt*
- *G. N. Yannakakis is with the Institute of Digital Games, University of Malta, Msida, Malta. Email: georgios.yannakakis@um.edu.mt*
- *J. Hallam is with the Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark. Email: john@mmmi.sdu.dk*

[14]. Such effects, however, are not trivial to show through a data modelling approach, since the objective ground truth is fundamentally ill-defined when ratings are treated as numerical values [5]. Therefore, the performance comparison between a regression and preference-learned model is irrelevant as the former is arguably *a priori* incapable of capturing the underlying affective phenomenon as precisely as the latter.

In summary, the key hypothesis this paper attempts to validate empirically is whether rank-based transformations of rating annotations yield more reliable models of affect than class-based transformations. To test this hypothesis we first create a synthetic affect model (testing "*in vitro*") from which we extract classes and ordinal rating annotations; these annotations are used to train CL and PL models that are compared to the ratings generated by the synthetic model. This gives us complete control over the input data and access to ground truth for measuring the accuracy of the obtained models. Second, we transform two real datasets annotated with ratings to a set of binary-class and pairwise preference datasets to test the algorithms "*in vivo*". The first dataset is a game-based physiological dataset annotated with ratings of seven affective and cognitive states. The second dataset consists of speech segments extracted from videos and annotated sequentially and continuously for arousal and valence. Various models of affect are trained using artificial neural networks on the binary classes and the pairwise preferences.

The analysis of the prediction performance across all datasets shows that transforming ratings into classes complicates the learning problem which suggests that the affect relations hidden in the data are being altered. Furthermore, the comparison between CL and PL models over the original datasets suggests that PL methods lead to more efficient, generic and robust models which capture more information about the ground truth (i.e. annotated affect). In addition, the analysis of the synthetic datasets clearly shows that PL models better approximate the underlying function between input (i.e. affect manifestations) and output (i.e. affect annotations).

The paper is structured as follows. Section 2 reviews the literature on learning from data provided in a rating format and Section 3 details the methodology used for our comparative studies. In Section 4 we describe the three datasets used to evaluate our hypothesis and Section 5 presents the results obtained. Finally, Section 6 presents the key findings and limitations of the approach followed and concludes the paper.

2 Background: Learning from Ratings

The task of affect modelling [18] consists of finding a function that maps a set of measurable inputs (e.g. extracted features from physiology) to a particular affective state. This is achieved by machine learning (i.e. automatically adjusting) the parameters of a model to fit a dataset that contains a set of input samples, each one paired with target outputs (i.e. supervised learning). The input samples correspond to the list of measurable attributes (or features) while the target outputs correspond to the annotations of affect for each of the input samples. The representation of the affect annotation determines the output of the model and, in turn, the type of the machine learning approach that can be applied. The three machine learning alternatives for learning from rating-based affect annotations, namely *regression*, *classification* and *preference learning*, are discussed in detail in this section.

2.1 Regression

When the outputs are represented as real-values that the model needs to approximate, the modelling problem is known as (metric or standard) regression. Ratings naturally define an ordinal scale [16], and while it is mathematically possible to use regression algorithms to learn the exact numeric ratings annotated by users (or experts), in general it should be avoided because regression methods transform implicitly the ordinal scale into a numerical scale (ratio scale). This transformation introduces two strong biases that we denote as *non-linear scale* and *subjectivity of ratings*.

- **Non-linear scale:** Even when ratings are given as numbers, the underlying subjective scale is not linear, i.e. the difference among questionnaire items is non-uniform [5]. For instance, in a 7-point scale of arousal the difference between 6 and 7 may be larger than the difference between 4 and 5 as some annotators rarely use the extremes of the scale or tend to use one extreme more than the other [14].
- **Subjectivity of ratings:** The difference among questionnaire items may change across several sessions of the same annotator (memory effects) and vary across different participants (e.g. because of cross-cultural differences [19] and the person-dependent internal scale [13]). In addition, equal ratings are considered to be exactly the same which may not be true as the questionnaires may not provide enough granularity (in addition to the memory effects already mentioned).
  This bias is typically minimised by gathering several reports from the same user and normalising the responses with the minimum and maximum values used (see [15] among others) which, in turn, exacerbates the non-linear scale bias and adds other types of experimental bias as the scale is artificially transformed.

For these reasons, practices that involve treating ratings as real-valued numbers (e.g. averaging ratings across annotators [20]) are fundamentally flawed. Prediction models trained to approximate a real-value representation of a rating — even though they may achieve high prediction accuracies — do not necessarily capture the true affect manifestations because the ground truth used for training and validation of the model has been undermined by the numerous biases discussed above.

Regression methods are popular within studies of continuous affect annotations (e.g. [21]). Even though annotations are given as sequences of real-values, one must bear in mind that the aforementioned limitations are still present in continuous annotations, and therefore, treating annotations as numerical instead of ordinal scales is a mathematically unsupported practice [5].

That said, there exist a few uses of regression in affect modelling that are methodologically sound with respect to the treatment of human reports. In particular, regression was used in [22] to learn the probability of a particular pose representing a discrete emotion. The probability was given as the number of annotators that tagged the posture with an emotion over the total number of annotations on it. Note that in that study the annotators used nominal reports (one item is selected from a list of emotions) instead of ratings; would the reports consist of ratings, a target probability must not be computed by averaging the ratings as that approach would introduce the biases described above.

The evaluation of regression methods is outside the scope of this paper. Further, we argue that the known fundamental psychological and psychometric pitfalls in human reports of affect described above provide sufficient evidence against the use of regression in affect modelling [12], [13], [14].

2.2 Classification

The inadequacy of regression for learning user ratings is mitigated by classification methods, which are commonly used instead. These methods expect a nominal value as a learning target — i.e. a value from a finite and non-structured set (of classes) — thus, the ordinal scale defining ratings needs to be transformed into a nominal scale. The common practice in affective computing consists of transforming sets of consecutive ratings into separate classes (e.g. see [7], [8], [15], [23] among many). As an example (see [9]), arousal ratings on a 7-point scale are transformed into *high*, *neutral* and *low* arousal classes using 7-5, 4 and 3-1 ratings, respectively. While this application of classification methods appears appropriate, the ordinal relation among classes is not being taken into account. More importantly, the transformation process adds to the subjectivity of ratings bias a new type of bias described below: *class splitting criteria*.

- **Class splitting criteria:** As the split criteria that create the data classes are designed by the machine learning/affective computing researcher, data artifacts are evidently generated. In other words, the dataset analysis is skewed by information that was not initially included in the affect annotation process. Certain annotation tools may define clearly separated classes; for instance, a positive and negative class could be created from a valence Likert scale with items 'Very negative', 'Negative', 'Neutral', 'Positive' and 'Very positive' with a minor bias (if the neutral responses are dropped). However, scales without clear class split points (e.g. self-assessment manikins for arousal) cannot trivially be split, adding inaccurate information to the representation of the affect annotations. This type of data bias is augmented when the class boundaries are not selected based on the physical meaning of the scale but based on requirements of the modelling method used, e.g. by splitting data to a balanced number of input samples per class. Such practices may lead to higher prediction accuracies but — as the derived classes are arbitrary from the annotator's perspective — the learned models may deviate further from the goal of capturing the true relationship between annotated affect and affect manifestation.

The non-linear scale bias is also existent if ratings from several annotators are averaged before creating the classes.

It is worth noting that classification is perfectly suited to the task when nominal annotations are provided instead of ratings, i.e. annotators select the emotion that is felt or displayed from a list of possibilities (see [24], [25], [26] among others). Note, however, that this experimental protocol may produce datasets with an unbalanced number of samples per class (which is required by most classification algorithms) and it does not provide information about the intensity of each emotion.

2.3 Preference Learning

As an alternative to regression and classification methods, preference learning methods are designed to learn ordinal relations. As ratings, by definition, express ordinal scales they can directly be transposed to any ordinal representation (e.g. pairwise preferences). For instance, given a participant's rating report indicating that a condition A felt 'slightly frustrating' and a condition B felt 'very frustrating', a PL method will train a model that predicts a higher level of frustration for B than for A, avoiding introducing the non-linear scale bias when assuming a fixed difference between 'very' and 'slightly'. The problem of the scale varying across time due to episodic memory still persists but can be minimised by transforming only consecutive reports, i.e. given a report for three conditions A, B and C, the model can be trained using only the relation between A and B, and B and C (not considering the comparison between A and C). The limitation of different subjective scales across users can be safely bypassed by transforming the affect ratings into ordinal relations on a per-subject basis. If conflicting annotations exist, a basic preference learning algorithm will favour the most common relations and the information of an even number of conflicting training samples would be neutralised. Particular training algorithms also allow the experimenter to introduce the confidence of a particular annotation that could also be used to favour some annotators over others (e.g. [27]). Finally, the class-split bias is eliminated altogether as this step is not required. An inherent requisite for these methods is that each annotator needs to provide at least two reports which is, however, common practice in affect annotation.

PL has already been successfully applied to learn affect annotations. Martínez et al. [28], [29] and Yannakakis et al. [30], [31] have extensively explored several approaches based on artificial neural networks to learn dissimilar affective and cognitive states reported as pairwise preferences. Garbarino et al. [32] have used *linear discriminant analysis* to learn pairwise enjoyment predictors. While not applied to affect modelling, it is worth noting that preference learning methods based on other popular classification and regression techniques have also been adapted to learn human subjective reports: e.g. the *ranking support vector machine* of Joachims [33] and the *Gaussian processes for preferences* of Chu [34]. In all these studies, preferences are used as the raw data and no comparison between preference learning and classification is provided.

Crammer and Signer [35] compared single-layer perceptron classification, regression and preference learning training algorithms in a task to learn exact ratings. They reported that the PL method outperforms the other methods on several synthetic datasets and a movie-ratings dataset. This paper is not concerned with the problem of learning exact ratings; instead, we focus on the comparison of PL methods against methods typically used in affective computing research that transform affect ratings into nominal representations.

3 Method

To test the main hypothesis of this paper, we construct and compare prediction models for ratings of affect using classification (CL) and preference learning (PL) in three datasets: one synthetic dataset and two known affective datasets (details about the three datasets can be found in Section 4).

As depicted in the general methodology followed in Fig. 1, we first partition the dataset into a validation set and a training set using a standard validation procedure (e.g. 3-fold cross-validation). Then, we apply a number of transformations to the training fold (see Section 3.1) to create CL and PL datasets. Note that these datasets do not contain ratings but their transformations to either binary classes or pairwise preferences; consequently, the size of the resulting training datasets is not necessary the same between the two methods. While the exact same information about the original ratings is available for both approaches, one could argue that the different training sizes generate an unfair comparison. To eliminate this potential effect, each experiment is repeated with a number of alternative transformations that create a large variety of sizes for both modelling approaches.

The modelling methodology followed for both approaches is based on backpropagation training of artificial neural networks (see Section 3.2). As the performance of this method can be affected by unbalanced data, we apply oversampling [36] to every training dataset. For classification, a balanced training dataset contains the same number of samples for positive and negative classes whereas for preference learning, in a balanced dataset the number of training pairs in which the first sample is preferred is the same as the number of pairs in which the second sample is preferred (i.e. order effects are removed).

In the course of finding an appropriate measure for comparing the two methods fairly we are faced with the following challenge: a preference model naturally answers the question *is A better than B?* for any given A and B, while the classification model answers the question *which class does A belong to?* for any given A. As the two are clearly not equivalent standard performance measures that rely on transformed datasets (such as percentage of correctly classified samples or the F1-measure) are not appropriate measures of performance and comparison. There is obviously a need of a common ground for their comparison after the models are trained and that can be found on a validation dataset that contains ratings. We thus compare the models using the *Kendall's tau* coefficient (τ) [37] which can be applied to compare the orders of the original ratings and the predictions of CL and PL methods on the validation set — avoiding any further transformation of the data. In that way, both algorithms are evaluated on the same dataset for their ability to predict unseen ratings. More details about the performance measure used can be found in Section 3.3.

In addition, as the exact target function is known in the synthetic dataset, we can evaluate the precision to which the original model is learned for that dataset. To measure this, we normalise the outputs of the target and the learned models and calculate the sum of squared deviations. This measure indicates how different the outputs between the synthetic model and either CL or PL models are.

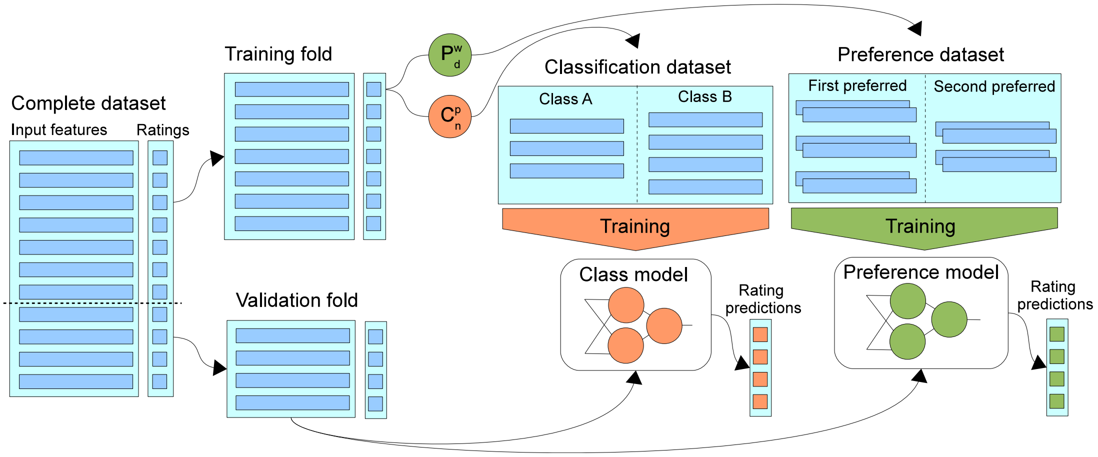


Fig. 1: Dataset preparation. The original dataset is partitioned into training and validation folds. To train a model using classification techniques, the ratings in the training fold are transformed into classes following the C_n^p constraints (see Sec. 3.1). Alternatively, to train a model using preference learning techniques the same ratings are transformed into pairwise preferences following the P_d^w constraints (see Sec. 3.1). The models trained following any of the two approaches (see Sec. 3.2) can be used to make a real-valued prediction of the ratings in the validation set. The performance measure used for comparing the two approaches is the *Kendall's tau* coefficient (τ) [37] between annotated and predicted ratings (see Sec. 3.3).

3.1 Transformation of ratings

This paper builds on the fundamental axiom that ratings follow ordinal scales and thus cannot correctly be treated either as numerical or as nominal values. To test this empirically we first transform available ratings into either nominal or ordinal formats depending on whether we are using classification or preference learning methods, respectively. This section describes the procedure followed for transforming the rating datasets into classification and preference learning datasets.

- **Classification** datasets are created by fixing a number of classes and assigning particular rating values to each class. For this study we use two classes (binary classification) because it yields single-output models which are directly comparable to models trained with PL. We select the ratings that fall within each class by setting two parameters: the positive (p) and the negative (n) class threshold which, respectively, specify the minimum rating value that falls within the positive class and the maximum rating value that falls within the negative class. Thus, data samples with ratings greater than or equal to p are assigned to the positive class, data samples with ratings lower than or equal to n are assigned to the negative class, and the remaining samples are discarded as neutral samples. For instance, if we transform a dataset containing 5-point Likert scale ratings using $p = 4$ and $n = 2$, all samples with rating equal to 4 or 5 are considered positive, all samples with rating equal to 1 or 2 are considered negative and all samples with rating equal to 3 are considered neutral.
- **Preference learning** datasets are assembled by converting the global order specified by ratings into partial pairwise orders. For every possible pair of rated input samples $(\mathbf{x^A}, \mathbf{x^B})$ within the same annotation session (e.g. same annotator on one day), we create a pairwise preference $(\mathbf{x^A} \succ \mathbf{x^B})$ in which the sample with a higher rating $(\mathbf{x^A})$ is preferred over the other sample $(\mathbf{x^B})$. For instance, if we transform a dataset containing 5-point Likert scale ratings, every sample rated with 5 is combined with every other sample with a lower rating (1 to 4) to create a pair where the sample rated with 5 is preferred; the same process is applied to samples rated with 4 (combined with samples rated between 1 and 3), and so forth. Partial orders with more than two elements could also be explored; however, most PL algorithms are designed to learn from pairs.

In this paper we use only pairs that compare samples within the same reporting session (e.g. the same participant) because it is expected that the absolute values of ratings will probably differ across participants and across sessions separated in time. In addition, we define two parameters, *memory window* and *minimum distance*, that allow us to eliminate potential effects introduced by reporting biases.

The memory window, denoted by w, defines the maximum number of consecutive ratings allowed to be compared. For instance, if $w = 1$ only consecutive ratings are used to create the pairwise preferences. On the other hand, if $w = \infty$

every possible pair is included in the training dataset. This parameter can control for biases on the data due to changes in the subjective scale across subsequent reports.

The minimum difference, denoted by d, specifies the minimum difference value between compared ratings that justifies a *clear* pairwise preference. All pairwise preferences which are not labeled as *clear* are discarded to reduce the uncertainty fed to the modelling algorithm. As an example, if we have four consecutive rated samples of arousal $\{A : -0.1, B : 0.3, C : 0.4, D : 0.0\}$ and we are using $w = 1$ and $d = 0.1$, our pairwise dataset contains the pairs $\{B \succ A\}$, $\{C \succ B\}$, and $\{C \succ D\}$, whilst if we increase the minimum difference to $m = 0.3$ only the pairs $\{B \succ A\}$ and $\{C \succ D\}$ are included. This parameter is mostly useful for continuous annotation tools (e.g. FeelTrace [38]) commonly used to annotate emotional dimensions such as arousal and valence in every frame of a video; a minimal variation on these continuous affect annotations may not convey relevant affect variations.

3.2 Affect Modelling Methods

In this section we describe the machine learning methods used for the comparison across variant representation schemes and for testing the hypotheses of this paper. The exact parameters used in our experiments are described in Section 5. Artificial neural networks are known to be universal approximators as they are able to approximate any continuous function [39]. In addition, artificial neural networks have already demonstrated their efficacy in affect modelling tasks [28], [29], [30], [40]. In this paper they are used to learn the mapping between affect manifestations (i.e. speech, physiology and gameplay content attributes) and a set of dissimilar cognitive and affective states and affective dimensions (i.e player experience states, arousal and valence). In particular, we use *backpropagation* [41] to train the connection weights of the artificial neural networks. This algorithm iteratively adjusts the value of the weights according to their contribution to a given error function. The topology of the network has to be selected manually or by using some form of systematic tuning. For binary **classification** we apply the most common error function, i.e. the sum of squared errors E_{SSE}, defined as follows:

$$E_{SSE}(f^{\mathbf{w}}, D) = \sum_{(\mathbf{x},y)\in D} \frac{1}{2}(y - f^{\mathbf{w}}(\mathbf{x}))^2 \qquad (1)$$

where D denotes the dataset used, $f^{\mathbf{w}}(\mathbf{x})$ is the output of the artificial neural network for the input sample $\mathbf{x}$ (between 0 and 1 when using a logistic sigmoid activation function) and y is its corresponding target output (0 and 1 for the negative and the positive affect class, respectively). The network implicitly learns to calculate the probability that a given input sample belongs to the positive class.

When ratings are transformed to **preferences** an alternative error function is required as no target output exists for each input sample. In this paper we use the regularised least squares E_{RLS} error function [27]:

$$E_{RLS}(f^{\mathbf{w}}, D) = \sum_{(\mathbf{x}^{\mathbf{P}},\mathbf{x}^{\mathbf{N}})\in D} \frac{1}{2}(1 - (f^{\mathbf{w}}(\mathbf{x}^{\mathbf{P}}) - f^{\mathbf{w}}(\mathbf{x}^{\mathbf{N}})))^2 \qquad (2)$$

where $\mathbf{x}^{\mathbf{P}}$ and $\mathbf{x}^{\mathbf{N}}$ represent a pair of training samples such that $\mathbf{x}^{\mathbf{P}}$ is preferred over $\mathbf{x}^{\mathbf{N}}$. For each training sample the output of the trained network represents a *utility* value which defines its order; for any two training samples, the one with the larger utility would be predicted as preferred. For instance, if the artificial neural network is trained on frustration preferences, the output of the artificial neural network is a predictor of the intensity of frustration, which is equivalent to the probability of being frustrated predicted by binary CL models.

The two error functions are similar with the difference that while E_{SSE} is minimised when outputs are close to the target output classes, E_{RLS} is minimised when artificial neural networks separate the outputs for training pairs as much as possible (i.e. when artificial neural networks maximise the margin between preferred and non preferred samples). If we consider the original ratings in a dataset, E_{SSE} will train models that approximate similar values for all the different ratings binned into the same class whilst E_{RLS} will train models that differentiate between each pair of ratings.

3.3 Performance Measure

As already mentioned, the comparison between PL and CL models is achieved through Kendall's tau (τ) coefficient [37] on the same validation set. The Kendall's tau defines a reliable and fair performance measure for comparing CL against PL as it is a measure of the correlation between two orders. The τ value can serve as a measure of the amount of information that one ordinal model has learned from an ordinal dataset which is equivalent to the amount of information that an artificial neural network has learned from a set of ratings in this paper.

In short, τ measures whether the artificial neural network model and the ratings define the same order over every pair of input samples within each session. We utilise the version of τ adjusted to account for ties [42] and we limit the calculation to each annotator and annotation session to avoid the assumption that the values of the ratings are absolute; that is, to account for personal differences and changes on the subjective perception of the rating scale along time.

4 Datasets

In this paper we evaluate the two alternative modelling methods on one synthetic and two dissimilar real datasets: the Maze-Ball (MB) dataset [30] and the Sensitive Artificial Listener (SAL) dataset [43]. This section details the core characteristics of each dataset examined.

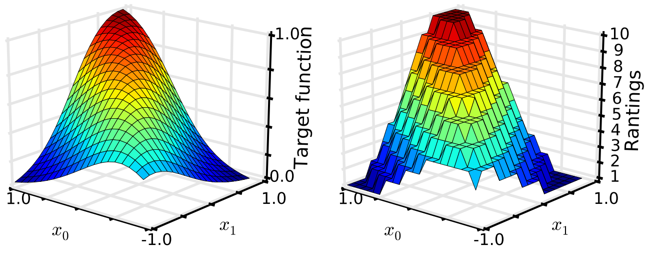
4.1 Synthetic Dataset

We create a synthetic dataset that contains a set of input attributes (features) and outputs in form of ratings. We use only two features for visualisation purposes as we are able to plot the target and learned functions and compare results by simple observation (see Fig. 2a). From this two-dimensional input space we sample a total of 200 samples within the $[-1, 1]$ interval. These input attributes only resemble two uniformly distributed features that are independent of each other regardless of their origin (e.g. one input could represent the average heart rate and the other could represent the minimum pitch interval between voice nuclei in a speech signal).

The features of each sample are used as inputs of a function whose output determines the rating of the sample. The function used to generate ratings for all 200 samples is an ad-hoc artificial neural network with one hidden layer containing two logistic sigmoid neurons. Many other topologies could have been tested but with a small hidden layer we maintain a sufficient level of function complexity without adding on the complexity of the artificial neural network topology. The weights of the network are selected randomly from the $[-1, 1]$ interval following a uniform distribution. In order to simulate personal biases in rating reporting, the 200 samples are randomly divided into 20 ordered groups; for each group a uniformly distributed random number that lies between 0 and 0.3 is generated and added to the output of the neural network. Each of the 20 groups represents a different artificial annotator, and the arbitrary order within the group simulates a sequence of ratings.

Finally, these outputs are used to yield two different types of rating on a 10-point scale as described below:

- Equidistant ratings: the range of the function ($[0, 1]$) is divided into 10 intervals of the same size (i.e. 0.1), each one assigned to a different rating item (see Fig 2b).
- Quadratic ratings: the range of the function ($[0, 1]$) is divided into 10 intervals, with the size of the intervals decreasing quadratically from central to extreme ratings. These ratings attempt to simulate reporting biases by which participants tend to use the extremes of the scale less often than the middle of the scale.



(a) Target function

(b) Equidistant ratings

Fig. 2: Synthetic dataset: target function and equidistant ratings for the two features considered (x_0, x_1).

4.2 Maze-Ball Dataset

The data of this set was gathered during an experimental game survey in which 36 participants played eight levels of the same video-game, i.e. *Maze-Ball*. The levels are selected from a pool of 8 different versions of the game that utilise alternative virtual camera behaviours that create different experiences. During the 90 seconds of each game, blood volume pulse and skin conductance were recorded at 31.25Hz using the IOM biofeedback device[1]. After each game, the players filled an online questionnaire reporting how the game felt with respect to *anxiety*, *boredom*, *challenge*, *excitement*, *fun*, *frustration* and *relaxation* on a 5-point Likert scale. The 5 Likert-scale alternatives for each of the above states were: *not at all*, *slightly*, *moderately*, *fairly* and *extremely*.

The physiological and gameplay information recorded for each game is transformed to a vector of real-valued features including average skin conductance, the average heart rate and the final score in the game. In total 10 features are used in this paper. The details of the Maze-Ball game design, the experimental protocol followed and the extracted features are already well reported in the literature and can be found in [30], [44].

4.3 SAL Dataset

The SAL corpus contains 739 1-second-long segments of speech extracted from 16 different videos. These videos feature the face of a person conversing with a virtual agent. Each segment is labeled for arousal and valence on a continuous scale ranging from -1 to 1 by several annotators using the Feeltrace software [38]. The ratings given by the most consistent annotator are used in this paper following the same procedure as in [43]. One may argue that this annotation does not feature an ordinal scale; note however, that the scale is still subjective and therefore not absolute (e.g., 0.3 does not stand for the exact same level of arousal every time it is used by the same or a different annotator) and the differences between numbers are

[1] www.wilddivine.com

not known (e.g. the actual difference of 0.2 between 0.8 and 1.0 of arousal is most likely not equal to the difference between 0.0 and 0.2 because the annotator might be scale-biased and, thus, not assigning the maximum value in the scale). Given these remarks, the most appropriate scale of measurement is ordinal [5].

Each video segment is defined by an input sample with 32 features including the minimum and average pitch interval between consecutive voiced nuclei. The details of the SAL dataset and the extracted features can be found in [43], [45].

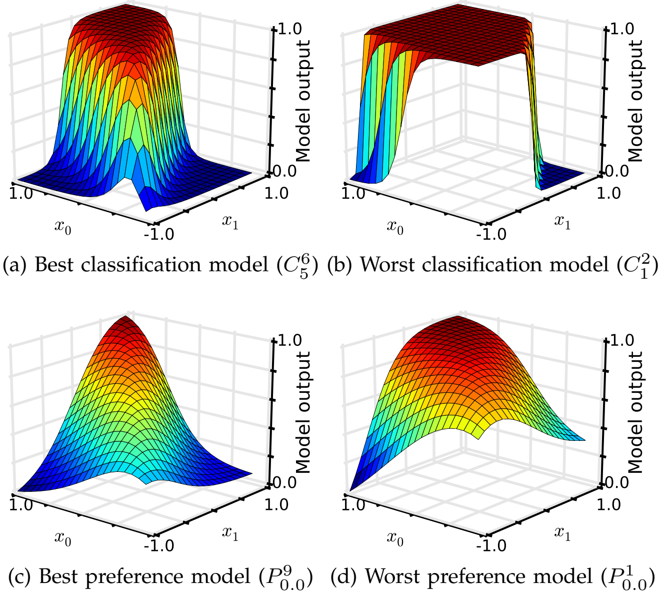
5 Experiments and Results

In this section we use the methodology presented in Sec. 3 to compare the performance of artificial neural network models of synthetic and affect datasets created using preference learning and classification methods.

Backpropagation training of the neural network models relies on a learning rate of 0.01 and 1000 epochs, after systematic parameter tuning experiments on the affect datasets. The classification experiments are performed using the machine learning tool Weka[2] and the preference learning experiments using the Preference Learning Toolbox[3]. This section details the experiments performed and the key findings of our analysis.

5.1 Experiments with Synthetic Datasets

For the experiments with synthetic data, we fix the topology of the neural networks trained to the exact topology of the synthetic model (2 logistic sigmoid hidden neurons). We then partition each set of the two rating types (equidistant and quadratic) into training and validation sets, resulting in sets of 130 and 70 samples respectively (corresponds to one iteration of 3-fold cross-validation). Then, we transform each training set into 9 classification datasets and 9 preference learning datasets. The different classification datasets are built by splitting the training samples into positive and negative classes using each pair of consecutive ratings as the threshold for the positive p and the negative n class (C_n^p for $n \in \{1, 2, ..., 9\}$ and $p = n + 1$). On the other hand, the preference datasets P_d^w are created with a fixed separation threshold ($d = 0.0$) and by varying the memory window from 1 to 9 ($P_{0.0}^w$ for $w \in \{1, 2, ..., 9\}$). The size of the training sets after oversampling varies from 136 to 258 samples in the classification sets and from 90 to 578 sample pairs in the preference sets. Note that, before oversampling, all the classification datasets contain the same number of samples whereas the number of samples in the preference learning datasets varies with the memory window size. We train 10 artificial neural networks for each dataset, and calculate the Kendall's tau between their outputs and the true ratings (see Fig. 4a and Fig. 5a).



(a) Best classification model (C_5^6) (b) Worst classification model (C_1^2)

(c) Best preference model ($P_{0.0}^9$) (d) Worst preference model ($P_{0.0}^1$)

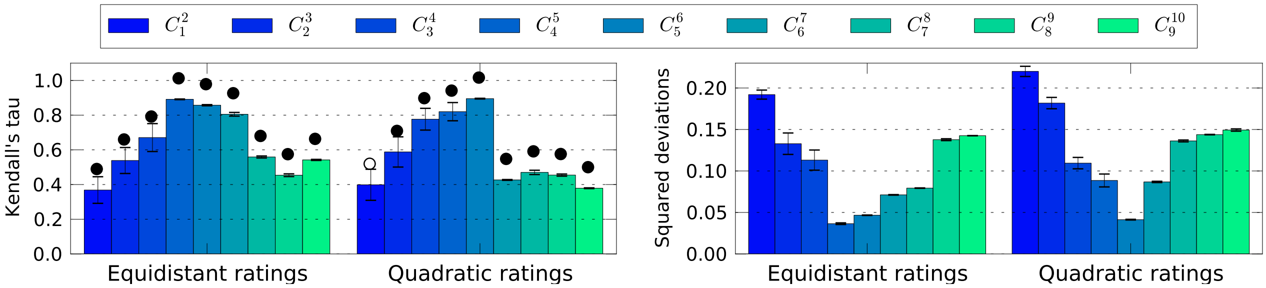
Fig. 3: Outputs of classification and preference learning models. Best and worst refer to the lowest and highest sum squared deviations values, respectively.

The most apparent difference between preference learning and classification is on their robustness: regardless of the transformation of ratings applied, preference learning yields models that predict the order of validation samples with high accuracy; on the other hand, classification only trains satisfactory models with particular transformations (specifically, C_4^5 and C_5^6). The majority of preference models reach τ values up to 0.9 while only classification models trained with the C_4^5 and C_5^6 yield τ values above 0.8.

The lowest performances for PL models are achieved with the smallest training dataset, $P_{0.0}^1$, which only contains pairs comparing consecutive samples; it appears that the number of pairs in that dataset is not sufficient to reconstruct the rating function as accurately as the other datasets.

For CL models, it is clear that when the class split point deviates from the middle of the scale, the performance starts dropping. Note that this result does not imply that the accuracy of CL models is low on the binary classification problem used for training. Instead, this finding points out that the original problem (i.e. learning the function underlying particular ratings depicted in Fig. 2a) has not been satisfactorily solved (i.e. the order of ratings in the validation set is not maintained). This is clear by observing the best and worst models learned from CL and PL training datasets (Fig. 3). Preference learning reconstructs smooth functions practically identical to

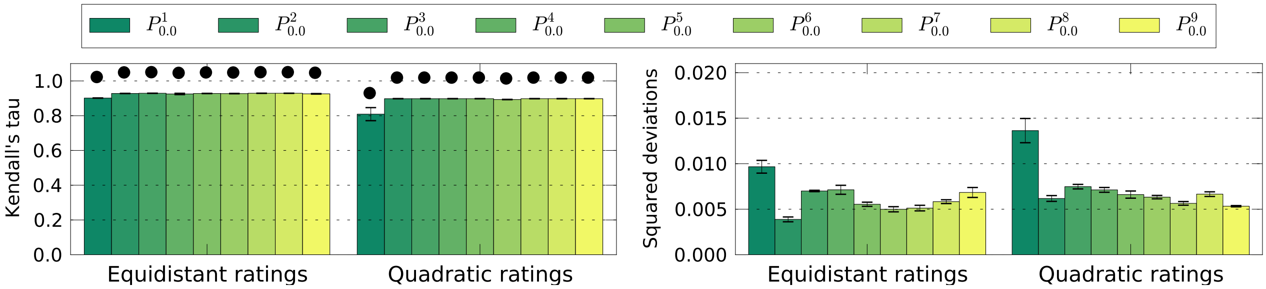
[2] Available at http://www.cs.waikato.ac.nz/ml/weka/
[3] Available at https://sourceforge.net/projects/pl-toolbox/



(a) Kendall's tau

(b) Sum of squared deviations

Fig. 4: Classification models for synthetic data: each bar represents the average Kendall's tau (a) or average sum of squared deviations (b) between the outputs of 10 models trained on one classification dataset C_n^p and the true ratings in the validation folds. Error bars depict the standard error and circles indicate a significant τ value (• represents $p - value < 0.05$ and ∘ represents $p - value < 0.1$).



(a) Kendall's tau

(b) Sum of squared deviations

Fig. 5: Preference models for synthetic data: each bar represents the average Kendall's tau (a) or average sum of squared deviations (b) between 10 models trained on one preference learning dataset P_d^w and the true ratings in the validation folds. Error bars depict the standard error and circles indicate a significant τ value (• represents $p - value < 0.05$ and ∘ represents $p - value < 0.1$).

the original function even with the worst model. On the other hand, classification only recreates a nearly-binary function that captures the largest differences. While these functions can predict accurately the binary classes, they are unable to capture the underlying ground truth. We can further reveal this pattern by analysing the sum of squared differences between the outputs of every trained model and the target function (see Fig. 4b and Fig. 5b). The enhanced robustness of preference models observed through Kendall's τ is also apparent in the difference between trained models and ground truth which are stable for PL models and highly variable across different CL datasets. Furthermore, note that the lowest difference for classification models (C_4^5 with difference of 0.036) is higher than the worst preference models ($P_{0.0}^1$ with difference of 0.013) as is already illustrated by the plots of the functions.

Overall, the results obtained on synthetic datasets suggest that preference learning methods are more reliable techniques than classification when applied to the task of learning predictors of ordinal variables such as ratings. More importantly, beyond producing models that approximate unseen ratings more accurately, PL models are closer to the target function (ground truth) than CL models. Consequently, it appears that preference learning is a more appropriate method, *"in vitro"*, to study the relationships between physical manifestations of affect and annotations given as ratings, as the learned mappings from one to another can be captured more accurately than when using classification methods. The key hypothesis of the paper remains to be tested in the next two sections where PL and CL are tested on real (*"in vivo"*) affective datasets.

5.2 Experiments with Affect Datasets

To further validate the results found with the synthetic datasets, we now apply both classification and preference learning to real affect datasets. We create a representative set of alternative classification and preference datasets from the ratings found in the two case studies, and use them to create models of affect which we compare.

We apply a standard 3-fold cross-validation procedure to partition each dataset into training and validation sets and run each experiment 10 times to minimise biases due to the random initialisation of the artificial neural networks. For each dataset we report the average and standard error (across the 30 models, 10 models per fold) of Kendall's tau. For each experiment several multi-layer perceptron topologies with

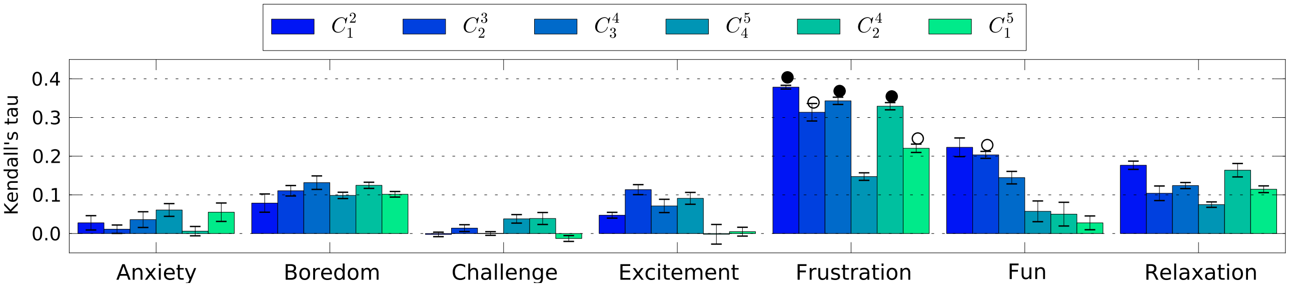


Fig. 6: Maze-Ball classification results: each bar represents the average Kendall's tau between 30 models trained on one classification dataset C_n^p and the true ratings in the validation folds. Error bars depict the standard error and circles indicate a significant τ value (● represents $p - value < 0.05$ and ○ represents $p - value < 0.1$).

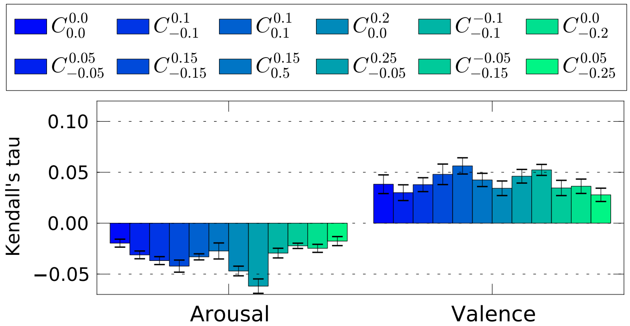


Fig. 7: SAL classification results: each bar represents the average Kendall's tau between 30 models trained on one classification dataset C_n^p and the true ratings in the validation folds. Error bars depict the standard error and circles indicate a significant τ value (● represents $p - value < 0.05$ and ○ $p - value < 0.1$).

no or one hidden layer were explored and we kept the one with highest prediction accuracy. This may yield optimistic results as a consequence of overfitting the validation folds, but as the same procedure is used for both approaches, the comparison between PL and CL is fair.

5.2.1 Classifying ratings

As it is not clear how to best transform ratings into classes (already demonstrated in the experiments of the synthetic dataset) we opt for generating binary-classes exhaustively and compare their potential for affect modelling. We thus split the input samples to two classes (i.e. whether the affective state is felt or not) following the procedure described in Section 3.1. In particular, the six *classification* datasets we have generated from MB ratings are as follows:

- The first four datasets are generated using each of the intermediate ratings of a 5-point Likert scale as threshold between the positive and negative class. The splitting points between 1 and 2 (C_1^2), between 2 and 3 (C_2^3), between 3 and 4 (C_3^4), and between 4 and 5 (C_4^5) define the border between the *not felt* and the *felt* class.
- The fifth dataset, C_2^4, ignores input samples rated at the midpoint of the 5-point rating scale (i.e. 3), and groups input samples with 1 and 2 rating values as the *not felt* class whereas it groups input samples with 4 and 5 rating values as the *felt* class.
- The sixth dataset, C_1^5, provides the highest possible separation margin between the two classes as it solely considers rating values of 1 and 5 as *not felt* and *felt*, respectively.

In addition to the MB datasets, we have generated twelve *classification* datasets from SAL ratings. To create a representative set of positive and negative boundaries, we vary the difference between the minimum rating in the positive class and the maximum rating in the negative class from 0.0 to 0.3 with 0.1 increments. In addition, we use three different values to centre that difference: -0.1, 0.0 and 0.1. For example, a difference of 0.0 centred at 0.0 ($C^{0.0}_{0.0}$) implies that every positive rating is assigned to the positive class and every negative rating is assigned to the negative class; on the other hand, a difference of 0.3 centred at 0.1 ($C^{0.25}_{-0.05}$) treats ratings above 0.25 as positive, ratings below -0.05 as negative and rating values in between as neutral (and thus discarded in this study).

The MB datasets C_1^2, C_2^3, C_3^4, and C_4^5 present the same number of samples across all affective states as all ratings are used. On the other hand, the number of samples in C_2^4 and in C_1^5 varies as a different number of them is removed for each state. Oversampling introduces variation on the size (but equal number of samples in each class) across datasets depending on the use of the 5-point scale on each affective state. For instance, the *boredom* dataset with the split C_1^2 kept almost the same size of the original dataset as approximately half of the reports indicated a rating value 1 (i.e. the game was not boring at all). On the other hand, some training folds for *anxiety* with splits C_4^5 and C_1^5 present few samples as most participants did not use the rating value 5 (i.e. no one felt extremely anxious during the game). C_2^3 yields the most balanced splits (and thus oversampled sets are more similar to the unbalanced sets) for *challenge*, *excitement*, *frustration*, *fun* and *relaxation*. Across all training sets the number of samples after oversampling varies from 32 to 348.

Likewise, the SAL datasets $C_{-0.1}^{-0.1}$, $C_{0.0}^{0.0}$ and $C_{0.1}^{0.1}$ contain the complete number of samples in each fold, which decrements as the interval of disregarded ratings increases. In this dataset oversampling not only presents significant variations across emotions and datasets but also across folds. For both affective dimensions, *arousal* and *valence*, $C_{0.1}^{0.1}$ yields the most balanced datasets (hence, oversampling has a minimal effect). Across all training sets the number of samples after oversampling varies from 370 to 744.

Figure 6 and Fig. 7 present the average τ values for the MB and SAL models, respectively, based on each of the different rating transformations. Most models trained on MB affective states estimate poorly the rating order in the validation fold; only *frustration* and *fun* models yield τ values above 0.2. C_1^2, C_3^4 and C_2^4 stand out for *frustration* and only C_1^2 and C_2^3 yield high correlations for *fun*. On the SAL dataset, none of the resulting models achieves correlations above 0.1. This is, in part, expected because of the large amount of information lost when transforming a continuous scale into a binary class. While for Maze-Ball this transformation groups together no more than 4 values, in SAL it groups a much larger number of distinct values which are then used in the calculation of τ.

Beyond the low expectations of an accurate model, it is somewhat surprising that none of the models trained for *arousal* yield positive correlations, suggesting that a general relation between speech features and annotated arousal does not exist in this dataset. Note that these values cannot be directly compared to the results reported in the literature because published studies on the SAL dataset evaluate the model performance based on nominal labels rather than ratings (e.g. see [43]). MB results are also not comparable to earlier studies with this dataset, as this is the first attempt to analyse rating annotations in this dataset (see [30], [44] for studies where rank annotations were used).

In summary, none of the attempts to convert ratings to binary classes generated models of affect that could accurately predict the order of ratings on unseen data. Furthermore, for the affective states that yield most accurate models, the particular split of the data used has a significant impact on the result, i.e. there is not a generic split that generates the most accurate models. In fact, the conclusion is obvious and quite the opposite: it still remains unclear how to best transform ratings into two classes even though it underpins the standard practice in affect annotation and affect modelling. This conclusion captures the key hypothesis of the paper that will be validated in the next section: what if one instead converts ratings to an ordinal scale like pairwise preferences and trains models on these datasets? Will the models be able to generalise better?

5.2.2 Ranking ratings

The *preference* datasets examined in this paper are created by comparing each rating within a session as described in Section 3.1. For MB we systematically vary the number of previous ratings — i.e. *memory window* — from 1 (for $P_{0.0}^1$ only comparisons between consecutive games are considered) to 7 (for $P_{0.0}^7$ every possible comparison is considered within a subject). In addition, as the scale is discrete we fix the minimum difference between ratings to the minimum difference (i.e. $d = 0.0$). After oversampling, the size of the training sets varies from 70 to 428 sample pairs.

The number of rated segments in a video of the SAL dataset is too large to explore every possible memory window, thus we select three representative values $w \in \{1, 3, 5\}$. In addition, as SAL ratings follow a continuous scale we also experiment with various minimum differences that determine if a difference between two ratings is a clear preference and thus, included in the pairwise dataset. We examine six indicative minimum differences values $d \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. As a result we examine a total of 12 sets after removing the combinations of memory windows and minimum difference that yield the most unbalanced training sets. The number of pairs in the selected sets varies between 48 and 2584 after oversampling is applied.

Increasing the number of compared ratings (i.e. memory window) provides a larger number of pairs; it may, however, add more reporting noise as a consequence of participants unconsciously varying their subjective rating scale during the full session (i.e. a maximum of eight 90-second-long games in MB and several minutes of watching video recordings in SAL). By limiting the number of comparisons we can reduce reporting inconsistencies as participants are required to provide ratings consistently only with fewer previous ratings. The minimum difference also has a significant effect on the number of pairs available. However, the pairs removed because they present a difference between ratings below the fixed minimum (d) are unlikely to provide relevant information given that very similar ratings in a continuous scale should not communicate a large difference in affect intensity, and they may in some cases add noise to the dataset (e.g. slight tremble in the hand of the annotator).

The experimental design of the MB game survey yields weak (if, in fact, really present) order effects and therefore oversampling increases slightly the amount of samples in the training datasets; imbalanced pairwise preference datasets generally reveal primacy or recency effects [30]. In addition, oversampling also replicates a small number of pairs for the PL datasets created from SAL, despite no measures were taken to minimise order effects in the experimental design.

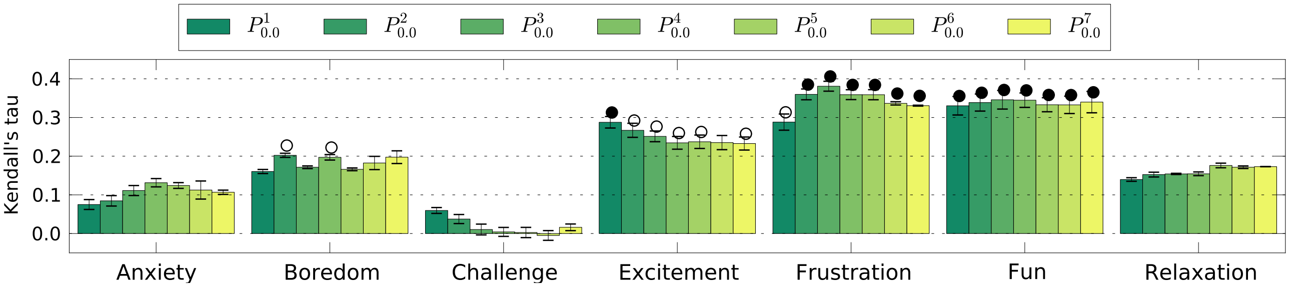


Fig. 8: Maze-Ball preference models: each bar represents the average Kendall's tau between 30 models trained on one preference learning dataset P^w_d and the true ratings in the validation folds. Error bars depict the standard error and circles indicate a significant τ value (• represents $p - value < 0.05$ and ∘ represents $p - value < 0.1$).

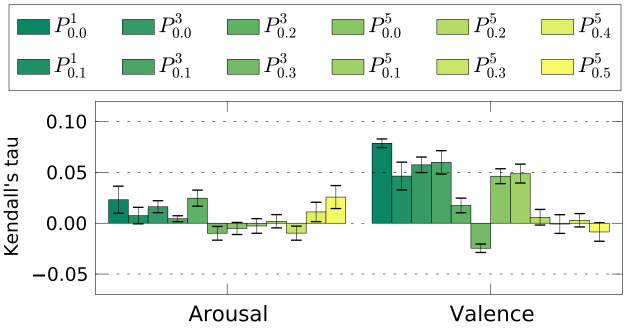


Fig. 9: SAL preference models: each bar represents the average Kendall's tau between 30 models trained on one preference learning dataset P^w_d and the true ratings in the validation folds. Error bars depict the standard error and circles indicate a significant τ value (• represents $p - value < 0.05$ and ∘ $p - value < 0.1$).

Figure 8 and Fig. 9 present the average τ values for the models based on each of the different rating transformations. Most datasets created for *frustration* and *fun* yield high order correlations with significant τ values over 0.3. Compared to the CL results, PL models not only reach higher values but are also more consistent and robust across data transformations. A similar pattern is observed for *excitement* datasets but with lower correlations overall (yet still statistically significant). For the remaining MB states, the correlations are in general as low as for the CL models. In MB, preference learning manages to yield significant τ values in the majority of experiments with *excitement*, *fun* and *frustration* whereas CL was only able to deliver significant effects for *frustration* and a single experiment on the *fun* state. As expected, the differences between PL and CL are not as notable as with the synthetic dataset. This is, in part, because in the MB dataset the function that relates input features to affect is not universal. In addition, the synthetic dataset uses a more fine-grained rating scale (10-point) that leads to a larger information loss when creating a binary classification set.

In the SAL dataset we observe similar results to the CL findings. Preference learning, similarly to CL, is not able to yield models of high performance value and the robustness of the two methods does not appear to be different either. In particular, most PL models yield similar performances to the CL models. Results on the SAL dataset suggest that neither method is able to learn from (and generalise over) the dataset which raises questions about the appropriateness of the dataset for either classification or preference learning. It may also signify that the lack of a temporal/dynamic dimension in the modelling approach undermines the performance of the model in this particular dataset (as also, in part, demonstrated with the recurrent artificial neural networks approach in [43]).

6 Discussion and Conclusions

The key findings of this paper challenge the current state-of-practice in affect annotation and modelling, of treating ordinal-scaled affect annotations (i.e. ratings) as nominal (i.e. classes). In particular, we examined the impact of the representation of ratings on the accuracy of models built on those annotations of affect. As ratings define the most common affective computing approach to affect annotation (e.g. given by Likert scales, the Self-Assessment Manikin and the Geneva wheel) our main hypothesis is that transforming ratings into ordinal values (rankings) yields more generalisable affect models when compared to transforming the same ratings into nominal values (classes).

To validate this hypothesis we tested the accuracy of affect models built on various class-based and rank-based representations within three datasets. These datasets present different types of data (synthetic and real), with different levels of rating resolution varying from 5-point (Maze-Ball dataset) and 10-point (synthetic dataset) Likert scales to continuous annotation (Sensitive Artificial Listener dataset), modalities of model input (physiology and speech) and annotation schemes (post experience and real-time); the exhaustive exploration of all above combinations would have been desirable but outside the scope of this paper.

The key finding of the paper based on these representative problems in affective computing is that rank transformations of ratings yield more generic models compared to class transformations of ratings across all case studies examined. Preference learned and classification models perform equally in cases where neither of the approaches manages to predict unseen ratings well (such as the arousal and valence states in the SAL dataset). Potentially, more powerful modelling approaches that incorporate temporal aspects — such as recurrent neural networks — are expected to demonstrate the superiority of PL in such cases, but this remains to be tested in future experiments.

We expected that the use of more fine-grained rating scales than the ones used in a standard Likert dataset (e.g. 5- to 7-point scales) would reduce the effects of transformation to nominal classes. Results on the 10-scale rating synthetic dataset, however, do not support this expectation. Rather, this paper has shown that, independently of the level of resolution of the rating scale, if ratings are not treated as ranked preferences effects such as the subjectivity of ratings and reporting inconsistency will remain [16] and undermine the generalisability of the models. We plan to test our hypotheses across more and different datasets including various types of affect annotation and user input modalities. Furthermore, the effect (or lack thereof) of the type and distribution of input features into CL and PL models needs to be analysed. We trust, however, that the results presented here are already generic across affect modelling tasks, application domains and user input modalities as they reveal a key limitation of traditional practice in affective computing: i.e. mistreating rating annotations of affect which, in turn, yield unreliable affect models.

The key findings of the paper suggest that ratings should not be treated as classes but the results also imply that the post-processing step of transforming ratings (to either classes or ranks) could be deemed unnecessary. Evidence (e.g. from findings in [16], [13]) already suggests that rank-based affect annotation should be preferred to rating-based annotation for its ability to eliminate annotation biases (cultural, subjective, inconsistency, inter-rater etc.). This paper not only further supports those findings but also empirically validates the hypothesis that rank based models built on ratings are more generic than class based models. It is fair to assume that the performance of any affect preference modelling approach would increase if data were directly annotated as ranks, which should, therefore, be the preferred approach whenever the annotation protocol allows it.

Acknowledgments

The authors would like to thank all subjects that participated in the reported experiments. The work is supported, in part, by the EU-funded FP7 ICT ILearnRW project (project no: 318803). Experiments reported in this paper run at the computational facilities procured through the European Regional Development Fund, Project ERDF-080 'A Supercomputing Laboratory for the University of Malta'.

References

[1] T. Mitchell *et al.*, *Machine learning*. McGraw-Hill New York:, 1997.

[2] R. Likert, "A technique for the measurement of attitudes." *Archives of psychology*, 1932.

[3] J. Morris, "Observations: Sam: The self-assessment manikinan efficient cross-cultural measurement of emotional response," *Journal of advertising research*, vol. 35, no. 6, pp. 63–68, 1995.

[4] K. Scherer, "What are emotions? and how can they be measured?" *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.

[5] S. S. Stevens *et al.*, "On the theory of scales of measurement," 1946.

[6] J. Fürnkranz and E. Hüllermeier, *Preference learning*. Springer, 2010.

[7] N. Liu, E. Dellandréa, B. Tellez, and L. Chen, "Associating textual features with visual ones to improve affective image classification," in *Affective Computing and Intelligent Interaction*, vol. 6974. Springer, 2011, pp. 195–204.

[8] M. S. Hussain, R. A. Calvo, and P. A. Pour, "Hybrid fusion approach for detecting affects from multichannel physiology," in *Affective Computing and Intelligent Interaction*, vol. 6974. Springer, 2011, pp. 568–577.

[9] J. Healey, "Recording affect in the field: Towards methods and metrics for improving ground truth labels," in *Affective Computing and Intelligent Interaction*, vol. 6974. Springer, 2011, pp. 107–116.

[10] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10, pp. 787–800, 2007.

[11] D. Wu, T. D. Parsons, E. Mower, and S. Narayanan, "Speech emotion estimation in 3d space," in *Multimedia and Expo (ICME), 2010 IEEE International Conference on*. IEEE, 2010, pp. 737–742.

[12] S. Ovadia, "Ratings and rankings: Reconsidering the structure of values and their measurement," *International Journal of Social Research Methodology*, vol. 7, no. 5, pp. 403–414, 2004.

[13] A. Metallinou and S. Narayanan, "Annotation and Processing of Continuous Emotional Attributes: Challenges and Opportunities," in *Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE)*, 2013.

[14] G. Langley and H. Sheppeard, "The visual analogue scale: its use in pain measurement," *Rheumatology international*, vol. 5, no. 4, pp. 145–148, 1985.

[15] J. Hernandez, R. R. Morris, and R. W. Picard, "Call center stress recognition with person-specific models," in *Affective Computing and Intelligent Interaction*, vol. 6974. Springer, 2011, pp. 125–134.

[16] G. N. Yannakakis and J. Hallam, "Rating vs. preference: A comparative study of self-reporting," in *Affective Computing and Intelligent Interaction*, vol. 6974. Springer, 2011, pp. 437–446.

[17] Y.-H. Yang and H. H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 762–774, 2011.

[18] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *Affective Computing, IEEE Transactions on*, vol. 1, no. 1, pp. 18–37, 2010.

[19] I. Sneddon, G. McKeown, M. McRorie, and T. Vukicevic, "Cross-cultural patterns in dynamic ratings of positive and negative natural emotional behaviour," *PloS one*, vol. 6, no. 2, p. e14679, 2011.

[20] F. Weninger, J. Krajewski, A. Batliner, and B. Schuller, "The voice of leadership: Models and performances of automatic analysis in online speeches," *Affective Computing, IEEE Transactions on*, vol. 3, no. 4, pp. 496–508, 2012.

[21] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *Affective Computing, IEEE Transactions on*, vol. 2, no. 2, pp. 92–105, 2011.

[22] H. Meng, A. Kleinsmith, and N. Bianchi-Berthouze, "Multi-score learning for affect recognition: The case of body postures," in *Affective Computing and Intelligent Interaction*, vol. 6974. Springer, 2011, pp. 225–234.

[23] P. G. Georgiou, M. P. Black, A. C. Lammert, B. R. Baucom, and S. S. Narayanan, ""that's aggravating, very aggravating": Is it possible to classify behaviors in couple interactions using automatically derived lexical features?" in *Affective Computing and Intelligent Interaction*, vol. 6974. Springer, 2011, pp. 87–96.

[24] R. S. Baker, G. R. Moore, A. Z. Wagner, J. Kalka, A. Salvi, M. Karabinos, C. A. Ashe, and D. Yaron, "The dynamics between student affect and behavior occurring outside of educational software," in *Affective Computing and Intelligent Interaction*, vol. 6974. Springer, 2011, pp. 14–24.

[25] A. Kleinsmith and N. Bianchi-Berthouze, "Form as a cue in the automatic recognition of non-acted affective body expressions," in *Affective Computing and Intelligent Interaction*, vol. 6974. Springer, 2011, pp. 155–164.

[26] T. Sobol-Shikler and P. Robinson, "Classification of complex information: Inference of co-occurring affective states from their expressions in speech," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 7, pp. 1284–1297, 2010.

[27] T. Pahikkala, E. Tsivtsivadze, A. Airola, J. Järvinen, and J. Boberg, "An efficient algorithm for learning to rank from preference graphs," *Machine Learning*, vol. 75, no. 1, pp. 129–165, 2009.

[28] H. P. Martínez, Y. Bengio, and G. N. Yannakakis, "Learning deep physiological models of affect," *Computational Intelligence Magazine, IEEE*, vol. 9, no. 1, pp. 20–33, 2013.

[29] H. P. Martínez, M. Garbarino, and G. N. Yannakakis, "Generic physiological features as predictors of player experience," in *Affective Computing and Intelligent Interaction*, vol. 6974. Springer, 2011, pp. 267–276.

[30] G. N. Yannakakis, H. P. Martínez, and A. Jhala, "Towards affective camera control in games," *User Modeling and User-Adapted Interaction*, vol. 20, pp. 313–340, 2010.

[31] G. N. Yannakakis and J. Hallam, "Entertainment modeling through physiology in physical play," *International Journal of Human-Computer Studies*, vol. 66, no. 10, pp. 741–755, 2008.

[32] M. Garbarino, S. Tognetti, M. Matteucci, and A. Bonarini, "Learning general preference models from physiological responses in video games: How complex is it?" in *Affective Computing and Intelligent Interaction*, vol. 6974. Springer, 2011, pp. 517–526.

[33] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," *Machine learning: ECML-98*, pp. 137–142, 1998.

[34] W. Chu and Z. Ghahramani, "Preference learning with gaussian processes," in *Proceedings of the international conference on Machine learning (ICML)*, 2005, pp. 137–144.

[35] K. Crammer and Y. Singer, "Pranking with ranking," *Advances in neural information processing systems*, vol. 14, pp. 641–647, 2001.

[36] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets," *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, 2004.

[37] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[38] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder, "'feeltrace': An instrument for recording perceived emotion in real time," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[39] A. Kolmogorov, *On the Representation of Continuous Functions of Several Variables in the Form of Super Positions of Continuous Functions of One Variable and Additive Functions*. SLA Translations Center, 1963.

[40] J. Wagner, J. Kim, and E. André, "From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification," in *Multimedia and Expo, IEEE International Conference on*. IEEE, 2005, pp. 940–943.

[41] D. Rumelhart, *Backpropagation: theory, architectures, and applications*. Lawrence Erlbaum, 1995.

[42] A. Agresti, *Analysis of ordinal categorical data*. John Wiley & Sons, 2010, vol. 656.

[43] K. Karpouzis, G. Caridakis, R. Cowie, and E. Douglas-Cowie, "Induction, recording and recognition of natural emotions from facial expressions and speech prosody," *Journal on Multimodal User Interfaces*, pp. 1–12, 2013.

[44] H. P. Martínez and G. N. Yannakakis, "Genetic search feature selection for affective modeling: a case study on reported preferences," in *Proceedings of the 3rd international workshop on Affective interaction in natural environments*, 2010, pp. 15–20.

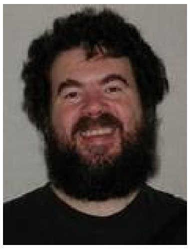
[45] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Raouzaiou, and K. Karpouzis, "Modeling naturalistic affective states via facial and vocal expressions recognition," in *Proceedings of the 8th international conference on Multimodal interfaces*, 2006, pp. 146–154.

Héctor P. Martínez is a postdoctoral researcher at the Institute of Digital Games, University of Malta. He obtained his Ph.D. degree from the IT University of Copenhagen in 2013 for his work on machine learning applied to player experience modelling. He received an MSc (2009) and BSc (2007) in Computer Science from the University of Valladolid.

His primary research interests include affective computing, multimodal interfaces, machine learning, game technology and serious games.

Georgios N. Yannakakis (S'04; M'05) is an Associate Professor at the Institute of Digital Games, University of Malta (UoM). He received the Ph.D. degree in Informatics from the University of Edinburgh in 2005. Prior to joining the Institute of Digital Games, UoM, in 2012 he was an Associate Professor at (and still being affiliated with) the Centre for Computer Games Research at the IT University of Copenhagen.

He does research at the crossroads of AI (computational intelligence, preference learning), affective computing (emotion detection, emotion annotation), advanced game technology (player experience modelling, procedural content generation, personalisation) and human-computer interaction (multimodal interaction, psychophysiology, user modelling). He has published over 150 journal and international conference papers in the aforementioned fields. He is an Associate Editor of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING and the IEEE TRANSACTIONS ON COMPUTATIONAL INTELLIGENCE AND AI IN GAMES.



John Hallam is currently Professor of Artificial Intelligence at the University of Southern Denmark (SDU), and Director of the Centre for BioRobotics at SDU. He graduated from the University of Oxford with an MA in Mathematics, and the University of Edinburgh with a Ph. D. in Artificial Intelligence.

His present research interests include biorobotics, evolutionary robotics, and applications of artificial intelligence and machine learning to affect modelling. He is President of the International Society for Adaptive behaviour.