

The Invariant Ground Truth of Affect

Konstantinos Makantasis, Kosmas Pinitas, Antonios Liapis, Georgios N. Yannakakis
Institute of Digital Games, University of Malta, Msida, Malta.

Email: {konstantinos.makantasis, kosmas.pinitas, antonios.liapis, georgios.yannakakis}@um.edu.mt

Abstract—Affective computing strives to unveil the unknown relationship between affect elicitation, manifestation of affect and affect annotations. The ground truth of affect, however, is predominately attributed to the affect labels which inadvertently include biases inherent to the subjective nature of emotion and its labeling. The response to such limitations is usually augmenting the dataset with more annotations per data point; however, this is not possible when we are interested in self-reports via first-person annotation. Moreover, outlier detection methods based on inter-annotator agreement only consider the annotations themselves and ignore the context and the corresponding affect manifestation. This paper reframes the ways one may obtain a reliable ground truth of affect by transferring aspects of causation theory to affective computing. In particular, we assume that the ground truth of affect can be found in the causal relationships between elicitation, manifestation and annotation that remain *invariant* across tasks and participants. To test our assumption we employ causation inspired methods for detecting outliers in affective corpora and building affect models that are robust across participants and tasks. We validate our methodology within the domain of digital games, with experimental results showing that it can successfully detect outliers and boost the accuracy of affect models. To the best of our knowledge, this study presents the first attempt to integrate causation tools in affective computing, making a crucial and decisive step towards general affect modeling.

Index Terms—causation theory, invariant features, outlier detection, affect modelling, games

I. INTRODUCTION

Affect modeling is largely viewed as a supervised learning task of subjectively defined labels. Due to labels' subjective nature, however, the ground truth of affect (which is attributed to such labels) is not easy to obtain reliably [1]. The dominant approach for creating consistent affective corpora is to treat the presence of subjectivity as a nuisance and to attempt to engineer it away by collecting ever larger datasets [2] or employing several independent annotators [3]. Although larger datasets may boost the performance of affect models and inter-annotator agreement methods can yield “clean” (i.e. consistent or reliable) labels, arguably both approaches aim to *objectify* the labels of emotions, which are by nature subjective.

In this study we embrace subjectivity as an inherent property of affect labels and introduce the notion of *causality* as a necessary element for reliable affect modeling. Inspired by causation theory [4] we assume that there exist features of affect manifestation that are *invariantly* related with affect labels. This implies that the relation that maps a subset of input features to affect labels remains invariant across different

users or tasks. We name the subset of features for which such a relation exists as *invariant features*. Our assumption is not new; on the contrary, every methodology for constructing handcrafted features within Affective Computing (AC) aims to transform affect manifestations to invariant features by implicitly assuming that such features exist. It is important to note that the notion of invariance we investigate in this paper is closely related to causation theory [5] where the existence of invariant features is a necessary—yet not sufficient—condition for the discovery of cause-effect relations.

To validate the aforementioned assumption, we introduce a method for automatically identifying the set of invariant features and use it to detect outliers; i.e. data points stemming from a specific annotator or task for which the relation that maps the invariant features to affect is different than the majority of data. Specifically, we consider as invariant those features which present similar correlation patterns with affect labels across different annotators and tasks. Thus, we create representations based on the computed correlation patterns, which reveal to which degree the data from different annotators or tasks exhibit a consistent relation between input features and affect labels. Finally, those representations serve as input to a conventional outlier detection algorithm, which brings together annotators and tasks that exhibit similar correlation patterns between input features and affect (defined as inliers), and, at the same time, identifies and discards data from outliers. After the outliers' removal, we use the identified invariant feature set to build robust models of affect.

Our method is of particular importance for AC corpora where disjoint sets of the data have been annotated by different annotators, such as first-person annotations. For this reason, we test our methodology within the domain of digital games using the first-person annotated AGAIN dataset [6]. Specifically we use data from fifty different participants that played three different games and self-annotated their gameplay in terms of arousal. Experimental results indicate that our method can effectively detect and remove outliers—i.e. inconsistent participants—resulting in robust models of affect. The boost in performance before and after outliers' removal is over 14% for one out of the three games and over 6.5% for the other two games. Notably, the performance of affect modeling for the set of outliers is on par with a random predictor.

This paper is novel in a number of ways. First, to the best of our knowledge, this study presents the first attempt to integrate causation tools in affect modeling; causation is expected to boost the generalization ability of affect models and arguably, generalizability is of utmost importance for any affective

This work has been supported by the EU Horizon 2020 research and innovation programme from the TAMED project (GA No. 101003397).

modeling method. Second, unlike any outlier detection method existent in AC, this paper proposes a methodology that detects outliers by considering both the affect measurements—i.e. the model’s input—and the subjectively defined affect labels—i.e. the model’s output. Third, the identified invariant features can be viewed as a causality-driven feature selection method that avoids capturing spurious correlations in affect modeling. Finally, we validate our approach extensively and across three different games played and annotated by 50 participants each. The validation results suggest that our methodology can efficiently identify and remove inconsistent participants resulting in consistent affective corpora and, thus, yielding more robust models of affect [5].

II. RELATED WORK

This section summarises related literature on techniques that aim to ensure consistent data. It reviews core outlier detection techniques used in machine learning and outlines methods employed to alleviate affect label ambiguities.

A. Outlier Detection in Machine Learning

Outlier detection is a fundamental data preprocessing task that aims to provide researchers with vital knowledge and assist them in making better decisions about their data. Outlier detection techniques strive to solve the problem of discovering patterns that do not adapt to expected behaviours [7]. In typical machine learning applications, deviation from the expected behaviour happens mainly due to noise in the data and failures of sensors that measure a phenomenon. Outlier detection techniques can be roughly classified into statistical, clustering, density, and learning-based methods [8].

Techniques based on statistical analysis use a stochastic distribution to model the data. Data points for which the probability to have been generated by the stochastic distribution model is below a threshold are denoted as outliers [9]. Some of the most commonly used statistical outlier detection methods are based on Gaussian Mixture Models [9], [10] and Kernel Density Estimation [11], [12].

Clustering approaches exploit clustering algorithms to describe data patterns and denote as outliers data points belonging to clusters with few members [13]. The effectiveness of such approaches is inherently dependent on the effectiveness of the employed clustering algorithm [14]. For detecting outliers via clustering one may employ partitioning algorithms such as K-Means [15] and CLARANS [16], hierarchical clustering algorithms [17], and density-based methods [18]. Clustering and outlier detection are two different tasks, however. The main drawback of clustering approaches, therefore, is the decision of whether a data point is an outlier or it represents a rare yet important aspect of the data.

Density-based methods consist one of the earliest known approaches to outlier detection, and they often serve as baselines for new algorithms [19], [20]. They assume that outliers lie in low-density areas of the data space, while inliers fall into high-density areas [21]. Although density-based approaches are entirely data-driven without making any assumption about

the distribution of the data, they are sensitive to parameter settings such as determining the neighbourhood size [22].

Finally, learning-based approaches operate in two stages. The first involves training a machine learning model to fit the data, and the second evaluates every data point against the trained model. A data point is denoted as an outlier when the model’s output deviates from a ground truth value [23]. Due to the recent success of deep learning, learning-based approaches have gained popularity. Deep learning methods, such as generative adversarial networks [24] and autoencoders [25], [26] are used to learn data representations which then are used with one-class classifiers [27] to detect outliers.

All the above techniques can be successfully applied to typical machine learning problems where the variables are objectively defined. They cannot be straightforwardly applied, however, to AC corpora due to the subjective nature of emotions. In this work, we embrace the subjectivity of affect labels and employ causation tools to formulate an outlier detection method tailored explicitly for affect corpora.

B. Consistent Labeling of Affect Datasets

Outlier detection techniques can be used as a data preprocessing step in AC to eliminate data points whose information is corrupted by noise or hardware failures. Beyond that, methodologies that attempt to ensure data consistency in AC primarily focus on the reliability of affect labels. More specifically, these methodologies compare multiple annotations of a particular context (e.g. image, video, sound) as provided by several independent annotators and test for annotation reliability in terms of inter-annotator agreement [3].

The simplest way to measure agreement is to count the number of items for which the annotators provide identical labels [28]. This agreement measure is highly expressive and easy to implement; it does not guarantee, however, a reliable annotation process as annotators may agree by mere chance. For this reason annotator agreement—which implies data reliability—is measured alternatively using coefficients from the kappa/alpha family, such as S [29], π [30], κ [31], [32] and α [33], which quantify agreement that was attained above the level expected by chance. Cronbach’s α has been used for instance in [34] to measure inter-annotator agreement and remove data points with unreliable labels. Yannakakis *et al.* [35] employ Krippendorff’s α to assess and compare the consistency between ordinal and interval affect labels. In [36], Cohen’s κ is used for ensuring consensus among annotators. Cronbach’s α and Cohen’s κ are used for assessing inter-annotator agreement in the RECOLA database [37], which has been used for audio/visual emotion recognition challenges.

Extensions for the above coefficients have also been proposed. For example, the study in [38] extends Cohen’s κ coefficient in scenarios where the annotators have the freedom to provide more than one label to a single data point. In [39] the signed differential agreement metric is proposed for measuring agreement in ordinal and interval-scale annotations, while in [40], [41] dynamic time warping is used to compare continuous annotation traces.

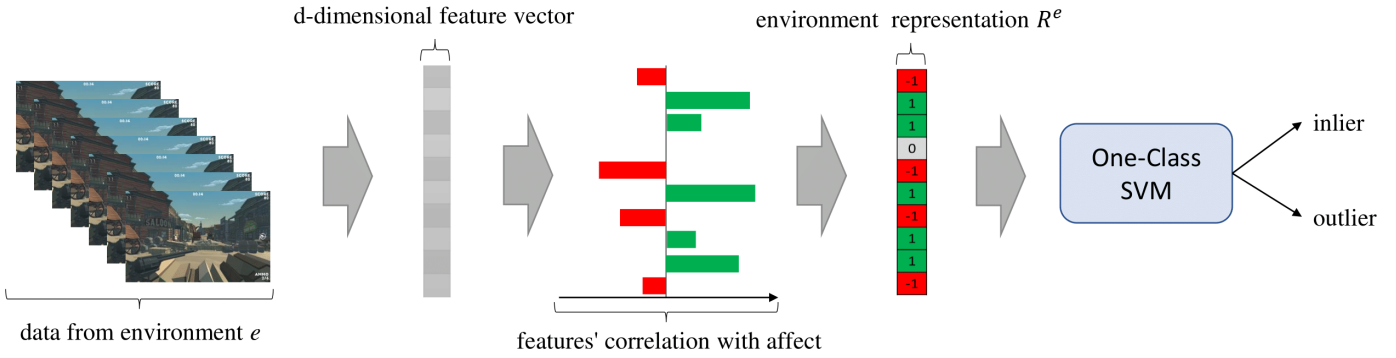


Fig. 1. The proposed overall methodology for detecting outlier environments.

The methodologies above require multiple annotations per data point to measure agreement and provide consistent labels. There are AC corpora however in which a) each affect elicitor (e.g. image) is labelled by a single annotator in e.g. a first-person manner and/or b) annotators have annotated different parts of it. The present study proposes a general approach to outlier detection for AC that bypasses the typical third-person multiple-annotation requirement and is applicable to any AC corpus for which different environments can be defined.

III. INVARIANT FEATURES FOR AFFECT MODELLING

This section presents our proposed methodology for detecting unreliable data points in affective corpora (III-A) and identifying invariant features for affect modelling (III-B). Both unreliable data detection and invariant feature identification are based on the notion of an *environment* that is used in the remainder of this paper. This term is borrowed from causation theory to refer to a mechanism that follows a specific probability distribution for generating data. In this study, we consider different participants as different environments since each participant follows an internal (personal) mechanism for generating data. Data in AC can be annotation traces (or labels) but also affect manifestations, and aspects of the affective interaction. It is important to note that a participant in a third-person annotation scheme is merely an annotator whereas a participant in a first-person annotation scheme can both annotate and generate data that can be annotated.

A. Removing Outliers among Annotators

Our approach is based on the assumption that there exist emotion elicitation mechanisms—encoded as stimulus-response rules—that are activated across different environments. Those rules can be represented by features that exhibit a similar relation to affect labels across different environments, i.e. invariant features. Based on the assumption above, we aim to discover those rules and discard data that violates them.

To define *invariance* we consider the correlation between features and their corresponding affect labels. More formally, let us define $\mathcal{D}^e = \{(X_i^e, Y_i^e)\}_{i=1}^{n^e}$ the data generated and annotated by environment $e \in \mathcal{S} = \{1, 2, \dots, E\}$. Affect measurements are represented by d -dimensional vectors $X_i^e \in \mathbb{R}^d$ and affect labels by scalars Y_i^e . For each environment e , we

compute the correlation between features X_i^e and labels Y_i^e . For continuous labels, the Pearson’s correlation coefficient can be used, while for binary labels the point-biserial correlation coefficient suits best. A positive (negative) correlation value between a feature x and affect labels indicates that the linear relation that maps x to labels is monotonically increasing (decreasing). Moreover, an undefined correlation value for a feature x implies that its variance equals zero (the value of x is constant for all data points in \mathcal{D}^e) and, thus, this feature does not carry any information about affect labels.

Based on the correlation coefficients, we represent the environment e by a d -dimensional vector $R^e \in \{-1, 0, 1\}^d$. Specifically, if the correlation between the i -th feature and labels is positive, then the i -th element of R^e equals 1, if the correlation is negative it equals -1, and if the correlation is undefined or zero it equals 0 (see Fig. 1). This d -dimensional environment representation encodes information about the way the different affect measurement features are related to affect labels. Therefore, environments with similar representations will share a large number of invariant features, that is, features that exhibit the same correlation patterns across similar environments. On the contrary, the set of invariant features for environments with dissimilar representations will be small. This implies that modelling affect, using machine learning models which are correlation-based, will be infected by spurious correlations.

We consider environments whose representation is dissimilar than the representations of the majority of environments as outliers. To detect which environments are outliers we use the One-Class SVM algorithm fed with environments’ representations R^e for $e = 1, 2, \dots, E$. One-Class SVM is an unsupervised learning algorithm that models expected data patterns to bring in one class all inliers and discard outliers. Fig.1 presents the outlier detection approach, which can be considered end-to-end since it receives as input a set of examples and outputs two clusters of data points: inliers and outliers.

B. Identifying Invariant Features

A critical byproduct of the methodology presented above is that the representation of the environments encodes information for relationships between features and affect labels that



Fig. 2. The three testbed shooter games of the AGAIN dataset [6] used in this study. From left to right: *Heist*, *Shootout*, and *Topdown*.

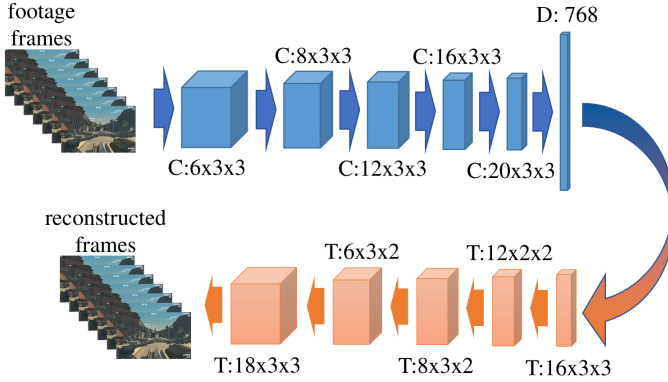


Fig. 3. The autoencoder architecture employed for constructing high-level representations for the visual information of gameplay time windows. The blue and the orange parts correspond to the encoder and decoder. “C”, “D” and “T” stand for convolutional, dense and transpose convolutional layers.

remain invariant across different environments. We exploit this information and use only those invariant features for affect modelling in an attempt to reduce spurious correlations even between similar environments. As mentioned above, the term environment corresponds to mechanisms that generate data by following different probability distributions. For a given affect corpus the number of environments is constant and different environments correspond to different data collection settings, such as different participants/annotators.

To identify invariant features, we follow a simple yet effective approach. Given a set of environments \mathcal{S} , we first discard all affect measurement features for which the correlation with affect labels is 0 or undefined. Then, we count the number of environments C_{pos} (C_{neg}) for which a feature is positively (negatively) correlated with affect labels. Finally, we denote a feature as invariant if:

$$\max(C_{pos}, C_{neg}) \geq \lambda |\mathcal{S}|, \quad (1)$$

where $|\mathcal{S}|$ is the number of environments in \mathcal{S} and $\lambda \in [0, 1]$ is a parameter for balancing between features’ invariance and modelling capacity. The value of λ is inversely proportional to the number of invariant features. For large values of λ the invariant features will not model spurious correlations, but due to their small number they may not exhibit enough modelling power to accurately predict affect. On the contrary, for small values of λ the invariance property is disregarded

resulting in models of affect infected by spurious correlations. We investigate the impact of parameter λ in Section V.

It is important to note that our method is *offline*, in the sense that it requires a batch of data to detect invariant features and inlier/outlier points. The detected invariant features, however, can be used in real-time affect modelling. For outlier samples, these features—like any set of features—will yield poor modelling results. For inlier points, however, the set of invariant features will improve modelling accuracy (see Section V).

IV. DATASET

This section presents the experimental protocol we designed for collecting the dataset used and the data preprocessing steps.

A. Experimental Protocol

To test our methodology we selected three games (*Heist*, *Shootout* and *Topdown*) from the AGAIN dataset [6] as depicted in Fig. 2. These games belong to the shooter genre and require accurate aiming and movement timing. Data was collected from 50 participants that were invited through Amazon’s mechanical Turk service. Each participant played a game with a duration of 120 seconds, and then annotated their recorded gameplay footage in terms of arousal using the RankTrace tool implemented in PAGAN [42], which allows continuous and unbounded annotations. This play-annotation cycle occurred for all three games. Since the participants had no prior experience in affect annotation, they were presented with an introductory screen that describes arousal as “*the intensity of gameplay no matter whether you like the game or not. High arousal can be a feeling of readiness, tension, excitement or exhilaration. Low arousal can be a feeling of fatigue, boredom, calmness or relaxation*”. Gameplay footage was recorded at 24 frames per second (24Hz) throughout the duration of the game. After play, participants annotated their gameplay video with an arousal trace, providing one arousal value for each of the recorded gameplay frames.

B. Data Preprocessing

To prepare the dataset for testing our methodology, we split the gameplay videos using non overlapping time windows of 3 seconds. Since RankTrace provides continuous and unbounded annotations, we first normalize the arousal annotations to a $[0, 1]$ value range in a game session-wise manner. Then, each window is assigned the mean arousal value of its frames. After

this data processing step, each window is described by the visual information of 72 frames and a single arousal value.

To create high-level representations for the visual information of the 3-seconds time windows, we use convolutional autoencoders. Specifically, we train one convolutional autoencoder with 5 convolutional layers, one dense layer and 5 transposed convolutional layers for each one of the games. For the first 4 convolutional layers we set the stride parameter to 2 and for the fifth layer we set it to 1. For all transposed convolutional layers we set stride to 2 and output padding parameter to 1. We use ReLU as the activation function and we train the autoencoders by minimizing the mean squared reconstruction error using the Adam optimizer. After training, we use the encoding part of the autoencoder to construct 768-dimensional representations for each gameplay time window. To reduce the computational cost of training, we convert the RGB frames to grayscale, fix their resolution to 240×150 pixels, and use frame skipping of 3 frames. As a result, the autoencoders’ input consists of 18 grayscale frames concatenated along the channels’ dimension. Details about the architecture of the autoencoder are presented in Fig. 3.

Having created high-level representations for the visual information of time windows, we form one dataset for each game for preference learning-based affect modelling [1]. On that basis, we identify all pairs of time windows belonging to the same gameplay. For each pair, we compare the arousal values of their corresponding time windows and we assign label 1 (0) when the arousal value of the first time window is larger (smaller) than that of the second window plus (minus) a threshold p_t . Finally, we disregard all pairs for which the absolute difference of their arousal values is smaller than p_t to avoid unstable preference learners due to trivial changes in their input. In this study, we set p_t to 0.15 which is slightly stricter than the threshold proposed in [6]. We represent every pair by a 768-dimensional latent vector, which corresponds to the representations difference of the first and second time window [43], [44]. Following the pairwise data transformation phase above, we end up with 50,254, 49,186 and 55,166 data points for Topdown, Heist and Shootout games, respectively.

V. RESULTS

In this section we present the experimental validation of our proposed methodology. We first present the results of the proposed outlier detection method (Section V-A) and we then present our affect modelling experiments using invariant features (Section V-B). For all performed experiments in this paper we evaluate the performance of our methodology in terms of preference prediction accuracy. Preference learning [45] can be seen as a binary classification task—preferred vs. non-preferred—where the two classes are perfectly balanced. Therefore, the prediction accuracy of a classifier that predicts the majority class in the training set is 0.5; this classifier serves as baseline. In this study we choose to use linear classifiers to be able to evaluate the performance of the proposed outlier detection method and affect modelling using invariant features

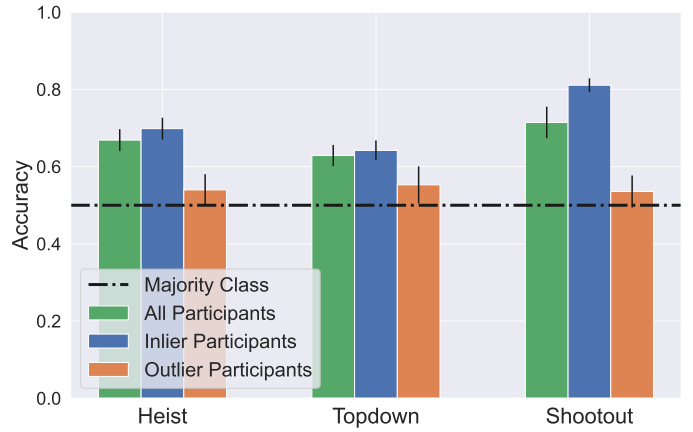


Fig. 4. Performance of the logistic regression affect model using all participants, the identified outlier participants, and the inlier participants.

without fiddling with complex nonlinear machine learning models, which add extra parameters for investigation.

Following the above preprocessing steps described in Section IV-B, we apply the outlier detection techniques on the time windows representations produced by our autoencoder. Moreover, since our affect labels are binary we use point-biserial correlation for creating the participants’ (environments) representations fed to the One-Class SVM. Finally, invariant features correspond to subsets of the 768-dimensional representations produced by the autoencoder.

A. Detecting Outlier Annotators

To investigate the effectiveness of the proposed outlier detection method, we first remove the outliers by applying the method presented in Section III-A, and then train three classifiers with the 768-dimensional representations obtained from our autoencoders. The first classifier uses data from all available participants without discriminating between inliers and outliers. The second and the third classifiers were trained using data from inlier and outlier participants, respectively. This way, we are able to evaluate modelling accuracy before and after outlier removal. To evaluate the performance of the classifiers we use the demanding leave-one-participant-out cross validation scheme and report classifiers’ preference prediction accuracy and 95% confidence intervals. Note that the classifiers are evaluated on different testing data. By using, however, classifiers of the same capacity and treating the training/testing sets as random variables we are able to objectively evaluate the quality of the outliers’ detection result (see [46] Theorem 3.3).

After the application of the proposed outlier detection technique, 16 participants are denoted as outliers for the Heist and Shootout games and 8 for the Topdown game. Fig. 4 presents the preference prediction accuracy and the 95% confidence intervals for the three classifiers mentioned above. The classifier that uses only inliers performs better (in the case of the Shootout game significantly better) than the classifier that uses data from all participants, and significantly better than the classifier trained only on outliers in all games. These

TABLE I
CLASSIFIERS’ ACCURACY AND 95% CONFIDENCE INTERVALS TRAINED ON OUTLIERS SET, RANDOMLY SELECTED INLIERS SETS AND RANDOMLY SELECTED PARTICIPANTS FROM THE POOL OF 50 PARTICIPANTS.

	Heist	Topdown	Shootout
Outliers set	0.54 ± 0.041	0.55 ± 0.048	0.53 ± 0.042
Random inliers sets	0.67 ± 0.010	0.62 ± 0.015	0.76 ± 0.006
Random part.	0.60 ± 0.009	0.60 ± 0.013	0.65 ± 0.016

findings suggest that the data from inlier participants is indeed consistent. Moreover, we can observe that the performance of the classifier that uses data from the set of outliers is slightly better than the performance of the baseline classifier.

To ensure that this drop in performance is not due to the size of the dataset (i.e. the outliers dataset is smaller than the inliers dataset), we run 30 independent experiments using data from randomly selected inliers such that the number of inliers is equal to the number of outliers; i.e. 16 randomly selected inliers for the Heist and Shootout games and 8 inliers for the Topdown game. These classifiers yield accuracies of 0.67, 0.62 and 0.76 for the Heist, Topdown and Shootout games, respectively (see Table I). These accuracies show a relative improvement of 24%, 13% and 43% compared to corresponding classifiers trained on the set of outliers. This indicates that the participants belonging to the set of outliers exhibit different patterns for playing and annotating the data. Due to this difference, outlier data are not consistent, resulting in affect models that are not able to capture the emotion dynamics. To benchmark the inlier detection algorithm against random chance, we also ran 30 independent experiments using data from randomly selected participants from the pool of 50 participants such that their number equals the number of outliers. Then, we train a classifier using their data (which may come both from the inlier and the outlier sets). These classifiers achieve accuracies of 0.60 for the Heist and Topdown games and 0.65 for the Shootout game, which have a relative drop of 11%, 9% and 22% compared to the accuracies of classifiers trained only on inliers sets with the same cardinality (see Table I). We can, therefore, argue that the outcome of the proposed outlier detection technique is not obtained by chance.

Based on all aforementioned results of this investigation, we can conclude that exploiting correlation patterns across different environments is an effective way to detect outliers.

B. Invariant Affect Modelling

In this section we investigate the impact of invariant features on arousal preference modelling. For that purpose we train classifiers using the feature sets obtained for different values of parameter λ in Eq. (1). We train classifiers for inlier and outlier participants and report their average accuracy and 95% confidence intervals following, once more, the leave-one-participant-out cross validation scheme.

The results of this investigation are summarized in Fig. 5. For all games the classifiers trained on inliers achieve the maximum accuracy for $\lambda = 0.7$. Specifically, these classifiers yield accuracy values of 0.71, 0.67 and 0.82 for the Heist,

Topdown and Shootout games, respectively. By setting the parameter λ to 0.7 the number of invariant features for the Heist game is 178, while for the Topdown and Shootout games are 120 and 130, respectively. We observe that our approach can reduce the number of features used for modelling arousal (initially 768) by identifying those features that are invariant across inlier participants and yield high modelling capacity.

The results obtained on the outliers set, however, do not seem to follow a specific pattern. Irrespective of the λ value, the average accuracy of the classifier is around 0.5, on par with random guessing. This suggests that the set of invariant features among the outliers that belong to the training set cannot be used for modelling the preferences of the outlier participant belonging to the test set. This highlights further the effectiveness of our proposed outlier detection technique; evidently the participants denoted as outliers all appear to follow inconsistent patterns between gaming (i.e. experiencing the context) and providing arousal annotations for their game.

By examining the bottom row of Fig. 5, we observe that the number of invariant features for the set of inliers is larger than the number of invariant features for the outliers. This observation is in alignment with the expected behaviour of the outlier detection scheme. The set of inliers comprises of data that are consistent: i.e. participants denoted as inliers follow similar patterns for playing and annotating the three games. Therefore, the probability the features of inlier data to exhibit similar correlation patterns with arousal labels is higher compared to outlier data. Due to this, the number of invariant features for the inlier data is larger than the corresponding number for the outlier participants. Interestingly, for the Shootout game there are no invariant features for the set of outliers for λ values larger than 0.7. We also observe that the accuracy of the classifier on the set of inlier participants drops slightly for λ values larger than 0.7. As discussed in Section III this behaviour is also expected. Parameter λ is used for balancing features’ invariance and modelling capacity. For λ values larger than 0.7 the selected features are invariant across the vast majority of participants; however, due their their small number they cannot accurately capture the arousal dynamics.

Based on the results above, we can conclude that employing invariant features—borrowed by causation theory—is an effective way towards deriving robust models of affect. Finally, parameter λ seems to affect the robustness of the models and, thus, it should be treated as a hyperparameter which should be appropriately set according to the affective corpus at hand.

VI. DISCUSSION AND CONCLUSIONS

In this paper we exploit invariant features, borrowed from causation theory, to derive an outlier detection method tailored for affective corpora and to build robust models of affect. Our proposed methodology is based on the assumption that there exist features of the experienced context (i.e. the affect elicitor) and the manifested affect that exhibit consistent relationships with annotated affect across different environments; environments are determined by different participants in this study. We identify those features via a correlation analysis and

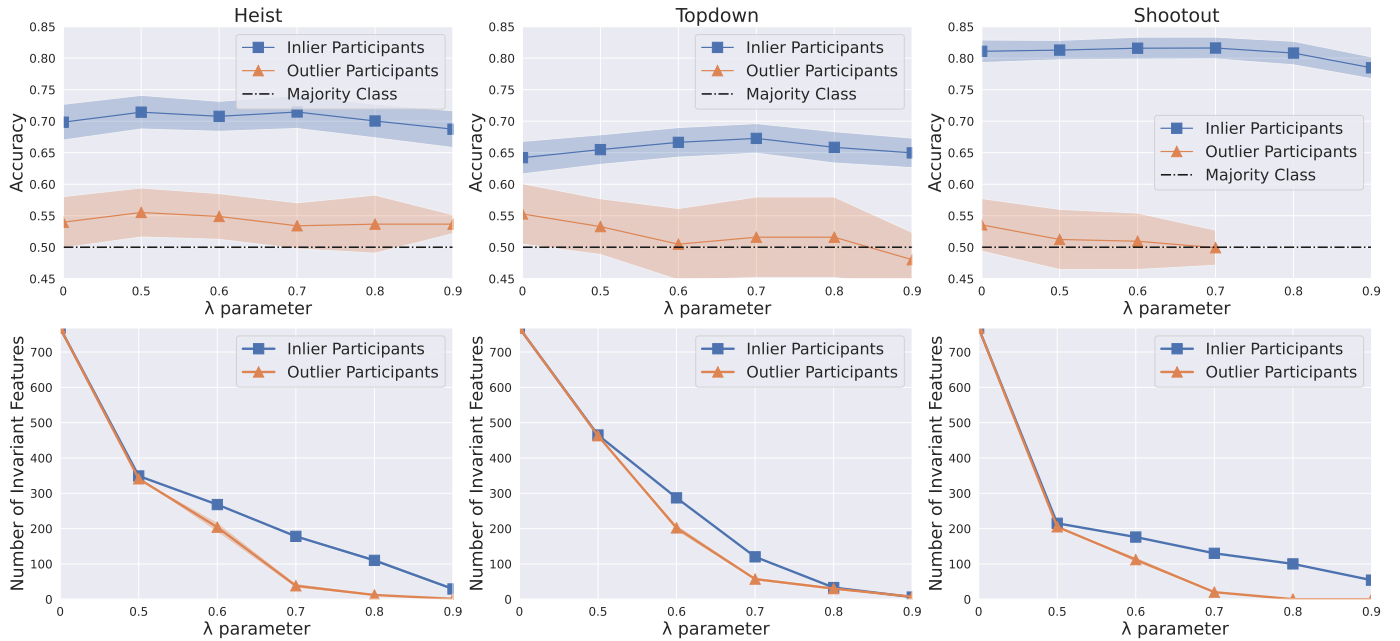


Fig. 5. Affect modelling accuracy and 95% confidence intervals (top row), and number of features used (bottom row) with respect to parameter λ .

we name them *invariant*. Based on the correlation patterns, we first create data representations for detecting outliers and then we use the set of invariant features to derive robust models of affect. We test our methodology within the domain of digital games by using data from three testbed games played and annotated in terms of arousal by 50 participants. The experimental findings suggest that our methodology can successfully denote as outliers those participants that present playing and annotation behaviour that is different than the majority of participants. Moreover, by using invariant features we are able to better capture arousal dynamics and build more accurate models of affect.

The presented methodology can be applied to any affective corpus comprising data from multiple environments. Its potential limitation, however, stems from the fact that the current implementation exploits linear correlation metrics to discover invariance. We thus need to further investigate the degree to which this linearity restricts the efficiency of our method. Towards this direction, we aim to employ non-linear correlation metrics and even learning-based approaches [47] for discovering invariant features. Moreover, besides linear classifiers, we aim to investigate the behaviour of invariant features in conjunction with non-linear models. In addition, this study focuses only on visual information. By using, however, general purpose representations of visual content, we do not restrict our method to specific tasks. Nevertheless, we aim to further validate its efficiency on data from diverse AC corpora that employ different and multiple modalities of affect. Finally, we aim to extend this approach by deriving several clusters (not only inliers/outliers) of data points that share common characteristics to avoid removing examples which makes the data less representative. Based on the clustering results we expect to be able to build cluster-wise models of

affect based on invariant features.

To conclude, to the best of our knowledge, this is the first attempt towards integrating methods from causation theory to affect modelling. Deriving models of affect that express causation is extremely important for the field. We believe that causation can provide valuable tools for disentangling complex relations between the context, its affect manifestations and corresponding affect labels and thereby identifying general cause-effect relationships or underlying rules of affect elicitation. At this point we should mention that some studies (e.g. [48]) propose the use of Granger causality [49] in affect modelling. This type of causality, however, tests the degree to which temporal information about affect measurements can forecast the experienced emotions. Therefore, Granger causality cannot discover cause-effect relations and, in general, it is not considered as a tool derived from causation theory. The proposed methodology has direct applications to any affect modelling task that employs data from multiple environments. In stark contrast to previously proposed methods for cleaning AC corpora, our method does not require multiple annotations per data point significantly reducing the data annotation cost.

REFERENCES

- [1] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Trans. of Affective Computing*, vol. 12, no. 1, pp. 16–35, 2018.
- [2] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. of Affective Computing*, vol. 10, pp. 18–31, 2017.
- [3] R. Artstein, "Inter-annotator agreement," in *Handbook of linguistic annotation*. Springer, 2017, pp. 297–313.
- [4] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proc. of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.
- [5] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. MIT Press, 2017.

- [6] D. Melhart, A. Liapis, and G. N. Yannakakis, "The Arousal video Game AnnotatIoN (AGAIN) dataset," *IEEE Trans. of Affective Computing*, 2022, accepted.
- [7] A. Ayadi, O. Ghorbel, A. M. Obeid, and M. Abid, "Outlier detection approaches for wireless sensor networks: A survey," *Computer Networks*, vol. 129, pp. 319–333, 2017.
- [8] S. Ranshous, S. Shen, D. Koutra, S. Harenberg, C. Faloutsos, and N. F. Samatova, "Anomaly detection in dynamic networks: a survey," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 7, no. 3, pp. 223–247, 2015.
- [9] K. Makantasis, A. Nikitakis, A. D. Doulamis, N. D. Doulamis, and I. Papaefstathiou, "Data-driven background subtraction algorithm for in-camera acceleration in thermal imagery," *IEEE Trans. of Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2090–2104, 2017.
- [10] X.-m. Tang, R.-x. Yuan, and J. Chen, "Outlier detection in energy disaggregation using subspace learning and gaussian mixture model," *International Journal of Control and Automation*, vol. 8, no. 8, pp. 161–170, 2015.
- [11] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos, "LocI: Fast outlier detection using the local correlation integral," in *Proc. of the IEEE International Conference on Data Engineering*, 2003, pp. 315–326.
- [12] L. J. Latecki, A. Lazarevic, and D. Pokrajac, "Outlier detection with kernel density functions," in *Proc. of the International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, 2007, pp. 61–75.
- [13] J. Zhang, "Advancements of outlier detection: A survey," *ICST Trans. of Scalable Information Systems*, vol. 13:1, pp. 1–26, 2013.
- [14] M. B. Al-Zoubi, "An effective clustering-based approach for outlier detection," *European Journal of Scientific Research*, vol. 28, no. 2, pp. 310–316, 2009.
- [15] A. Georgogiannis, "Robust k-means: a theoretical revisit," *Advances in Neural Information Processing Systems*, vol. 29, 2016.
- [16] R. T. Ng and J. Han, "Clarans: A method for clustering objects for spatial data mining," *IEEE Trans. of Knowledge and Data Engineering*, vol. 14, no. 5, pp. 1003–1016, 2002.
- [17] C. T. Zahn, "Graph-theoretical methods for detecting and describing gestalt clusters," *IEEE Trans. of computers*, vol. 100, no. 1, pp. 68–86, 1971.
- [18] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: Ordering points to identify the clustering structure," *ACM Sigmod record*, vol. 28, no. 2, pp. 49–60, 1999.
- [19] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," in *Proc. of the ACM SIGMOD inter. conf. on Management of data*, 2000, pp. 93–104.
- [20] W. Jin, A. K. Tung, J. Han, and W. Wang, "Ranking outliers using symmetric neighborhood relationship," in *Proc. of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2006, pp. 577–593.
- [21] E. Schubert, A. Zimek, and H.-P. Kriegel, "Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection," *Data mining and knowledge discovery*, vol. 28, no. 1, pp. 190–237, 2014.
- [22] H. Wang, M. J. Bah, and M. Hammad, "Progress in outlier detection techniques: A survey," *IEEE Access*, vol. 7, pp. 107 964–108 000, 2019.
- [23] P. I. Dalatu, A. Fitrianto, and A. Mustapha, "A comparative study of linear and nonlinear regression models for outlier detection," in *Proc. of the International Conference on Soft Computing and Data Mining*. Springer, 2016, pp. 316–326.
- [24] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.
- [25] E. Protopapadakis, A. Voulodimos, A. Doulamis, N. Doulamis, D. Dres, and M. Bimpas, "Stacked autoencoders for outlier detection in over-the-horizon radar signals," *Computational Intelligence and Neuroscience*, 2017.
- [26] S. Eduardo, A. Nazabal, C. K. Williams, and C. Sutton, "Robust variational autoencoders for outlier detection and repair of mixed-type data," in *Proc. of the International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 4056–4066.
- [27] P. Oza and V. M. Patel, "One-class convolutional neural network," *IEEE Signal Processing Letters*, vol. 26, no. 2, pp. 277–281, 2018.
- [28] P. S. Bayerl and K. I. Paul, "What determines inter-coder agreement in manual annotations? a meta-analytic investigation," *Computational Linguistics*, vol. 37, no. 4, pp. 699–725, 2011.
- [29] E. M. Bennett, R. Alpert, and A. Goldstein, "Communications through limited-response questioning," *Public Opinion Quarterly*, vol. 18, no. 3, pp. 303–308, 1954.
- [30] W. A. Scott, "Reliability of content analysis: The case of nominal scale coding," *Public opinion quarterly*, pp. 321–325, 1955.
- [31] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20:1, pp. 37–46, 1960.
- [32] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.
- [33] K. Krippendorff, *Content analysis: An introduction to its methodology*. Sage publications, 2018.
- [34] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PLoS one*, vol. 12, no. 3, p. e0173392, 2017.
- [35] G. N. Yannakakis and H. P. Martinez, "Grounding truth via ordinal annotation," in *Proc. of the international conference on affective computing and intelligent interaction*, 2015, pp. 574–580.
- [36] L. Devillers, R. Cowie, J.-C. Martin, E. Douglas-Cowie, S. Abrilian, and M. McRorie, "Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches," in *Proc. of the International Conference on Language Resources and Evaluation*, 2006.
- [37] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Proc. of the IEEE international conference and workshops on automatic face and gesture recognition*, 2013.
- [38] P. K. Bhowmick, A. Basu, and P. Mitra, "An agreement measure for determining inter-annotator reliability of human judgements on affective text," in *Proc. of the workshop on Human Judgements in Computational Linguistics*, 2008, pp. 58–65.
- [39] B. M. Booth and S. S. Narayanan, "Fifty shades of green: Towards a robust measure of inter-annotator agreement for continuous signals," in *Proc. of the International Conference on Multimodal Interaction*, 2020, pp. 204–212.
- [40] K. Makantasis, A. Liapis, and G. N. Yannakakis, "From pixels to affect: A study on games and player experience," in *Proc. of the International Conference on Affective Computing and Intelligent Interaction*, 2019.
- [41] —, "The pixels and sounds of emotion: General-purpose representations of arousal in games," *IEEE Trans. of Affective Computing*, 2021.
- [42] D. Melhart, A. Liapis, and G. N. Yannakakis, "PAGAN: Video affect annotation made easy," in *Proc. of the International Conference on Affective Computing and Intelligent Interaction*, 2019, pp. 130–136.
- [43] E. Camilleri, G. N. Yannakakis, D. Melhart, and A. Liapis, "PyPLT: Python preference learning toolbox," in *Proc. of the International Conference on Affective Computing and Intelligent Interaction*, 2019, pp. 102–108.
- [44] T. Joachims, "Optimizing search engines using clickthrough data," in *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 133–142.
- [45] J. Fürnkranz and E. Hüllermeier, "Preference learning," in *Encyclopedia of Machine Learning*. Springer, 2011, pp. 789–795.
- [46] J. Langford and R. Schapire, "Tutorial on practical prediction theory for classification," *Journal of machine learning research*, vol. 6, no. 3, 2005.
- [47] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv:1907.02893*, 2019.
- [48] T. Mittal, P. Mathur, A. Bera, and D. Manocha, "Affect2mm: Affective analysis of multimedia content using emotion causality," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5661–5671.
- [49] K. J. Friston, A. M. Bastos, A. Oswal, B. van Wijk, C. Richter, and V. Litvak, "Granger causality revisited," *Neuroimage*, vol. 101, pp. 796–808, 2014.