Nonlinear Appraisal Modeling: An Application of Machine Learning to the Study of Emotion Production

MEULEMAN, Ben, SCHERER, Klaus R.

Abstract
Appraisal theory of emotion claims that emotions are not caused by "raw" stimuli, as such, but by the subjective evaluation (appraisal) of those stimuli. Studies that analyzed this relation have been dominated by linear models of analysis. These methods are not ideally suited to examine a basic assumption of many appraisal theories, which is that appraisal criteria interact to differentiate emotions, and hence show nonlinear effects. Studies that did model interactions were either limited in scope or exclusively theory-driven simulation attempts. In the present study, we improve on these approaches using data-driven methods from the field of machine learning. We modeled a categorical emotion response as a function of 25 appraisal predictors, using a large dataset on recalled emotion experiences (5901 cases). A systematic comparison of machine learning models on these data supported the interactive nature of the appraisal–emotion relationship, with the best nonlinear model significantly outperforming the best linear model. The interaction structure was found to be moderately hierarchical. Strong main [...]
Nonlinear Appraisal Modeling: An Application of Machine Learning to the Study of Emotion Production

Ben Meuleman and Klaus R. Scherer

Abstract—Appraisal theory of emotion claims that emotions are not caused by “raw” stimuli, as such, but by the subjective evaluation (appraisal) of those stimuli. Studies that analyzed this relation have been dominated by linear models of analysis. These methods are not ideally suited to examine a basic assumption of many appraisal theories, which is that appraisal criteria interact to differentiate emotions, and hence show nonlinear effects. Studies that did model interactions were either limited in scope or exclusively theory-driven simulation attempts. In the present study, we improve on these approaches using data-driven methods from the field of machine learning. We modeled a categorical emotion response as a function of 25 appraisal predictors, using a large data set on recalled emotion experiences (5,901 cases). A systematic comparison of machine learning models on these data supported the interactive nature of the appraisal-emotion relationship, with the best nonlinear model significantly outperforming the best linear model. The interaction structure was found to be moderately hierarchical. Strong main effects of intrinsic valence and goal compatibility appraisal differentiated positive from negative emotions, while more specific emotions (e.g., pride, irritation, despair) were differentiated by interactions involving agency appraisal and norm appraisal.

Index Terms—Emotion, appraisal theory, machine learning, interactions, modeling

1 INTRODUCTION

1.1 Emotion Production

The question of what causes an emotion has been addressed by numerous theories in psychology. Prominent among these is appraisal theory, which claims that emotions are not caused by “raw” stimuli, as such, but by the subjective evaluation (appraisal) of those stimuli. For example, depending on whether you have the winning lottery number, the number drawn on television may either cause joy or disappointment. The number itself has no intrinsic emotional meaning, but has to be appraised with respect to its personal relevance and implications.

Appraisal theory subscribes to the view that emotions are multi-componential phenomena (Fig. 1), consisting of changes in appraisal, physiology (e.g., heart rate, blushing), expression (e.g., smiling, shouting), motivation (e.g., fight, flight), and subjective feeling [1], [2]. Each of these components contributes to the overall phenomenology of emotion (like symptoms of a syndrome), but the appraisal component takes central stage due to its presumed causal role in the emotion production process. Appraisal theorists believe that changes in the other four components are (largely) a consequence of changes in appraisal [1]. This causal hypothesis makes the appraisal-emotion relationship of special interest for the study of emotion production, and has generated much theoretical development, empirical research, and applications in affective computing [3]. By now, there are numerous studies that support the influence of appraisal on feeling, expression, physiology, and motivation [4].

Studying the influence of appraisal on emotion calls for explicit assumptions about the nature of that relation. Even a simple analysis requires a model that specifies what is input and output, and what kind of mapping algorithm connects input to output. So far, research has been mostly concerned with the former, investigating which appraisal variables differentiate a small number of basic emotions. Far less attention has been devoted to the choice of mapping algorithm. In the next section, we will argue step by step why this choice cannot be glossed over. First, we will discuss appraisal theories, highlighting interactional hypotheses which imply (potentially) complex mapping algorithms. Following this, we review how traditional approaches to appraisal-emotion modeling have handled the interaction assumption. These approaches can be broken down into two areas: theory-driven modeling and data-drive modeling. Theory-driven modeling is typically found in pure appraisal theory and the field of affective computing, where the express purpose is to create a computational model of emotion production. Data-driven modeling is typically found in psychological studies that collect data, and then apply a model to estimate relations between appraisal and emotion (e.g., linear regression). We will identify problems with the current practice in both of these approaches. Finally, we propose an improved data-driven strategy based on methods from the field of statistical machine learning, and show how this strategy can aid our understanding of emotion production.
One could question the main effect (circumstance, agent, or object), or whether an event elicits a different emotion depending on whether you yourself broke the vase (guilt) or your friend broke the vase (anger). In this scenario, appraisal of goal obstruction (the breaking of the vase) is modified by appraisal of agency (self versus other). In statistical terms, obstruction and agency are said to interact to differentiate emotions. An interaction between variables is a nonlinear effect, a point to which we shall return shortly.

Interactions in appraisal theories typically have a hierarchical structure, like a tree of decision rules. A main effect will first separate major clusters of emotions, followed by interactions that modify the main effect and differentiate increasingly specific emotions. A main effect in appraisal theories is often considered to be appraisal of relevance, determining whether an emotion is experienced at all [5]. Irrelevant events are not expected to elicit emotions. This effect is so basic, in fact, that appraisal theorists will sometimes omit it from prediction tables completely, instead focusing on the differentiation between emotions. Between emotions, a main differentiating effect is often considered to be appraisal of goal compatibility. This appraisal checks whether a stimulus or event is conducive or obstructive to one’s goals and desires, and is assumed to differentiate positive from negative emotions [4], [7]. For the present study, we will also consider goal-compatibility as the expected main effect, as our data do not contain non-emotional episodes.

What are the main effects and interactions predicted by specific appraisal theories? In Roseman’s model [7], appraisal of unexpectedness is assumed to separate surprise from all other emotions. Within the other emotions, appraisal of goal compatibility (consistent or inconsistent with one’s goals) is assumed to separate positive from negative emotions. Finally, specific differentiation is made by further appraisals of agency, coping potential and certainty. Thus, for a group of positive and negative emotions—as we have in our data—Roseman’s model predicts goal compatibility as the main effect, and agency, coping potential, and certainty as interactions. Lazarus also considered the interactive role of coping potential, distinguishing primary from secondary appraisal [6]. Primary appraisal refers to an evaluation of personal implication, and asks whether “something is at stake” for the person (relevance, novelty, and goal-compatibility). Conditional on this primary appraisal, a secondary appraisal concerns how the subject can cope with these implications, especially if goal obstructive [6]. The OCC model of Ortony et al. [12] proposes an explicit tree-structured chain of appraisals to differentiate emotions, with appraisal of agency as the main effect (circumstance, agent, or object), followed by interaction effects of goal compatibility, relevance, and even more agency (other versus self) to differentiate specific emotions. A hierarchical appraisal structure such as the OCC model implies that the main effect (i.e., the root of the decision tree) must have a non-zero value before appraisal values at the next level can influence emotion. Scherer’s Component Process Model (CPM; [8], [9], [10], [11]) makes this assumption the most explicit by assuming that appraisal takes place according to a hierarchical sequence of relevance, implication, coping potential, and normative significance checks. Here, appraisal of relevance is the main effect, determining whether an emotion occurs in the first place. Scherer’s prediction tables typically focus on between-emotion differentiation, however, for which appraisal of goal-compatibility is the main effect separating positive from negative emotions. Appraisals of coping potential and normative significance are assumed to interact with this appraisal to produce increasingly specific emotions (see [10] for a predicted table of appraisal-emotion configurations).

Theoretical appraisal models originated from common-sense thinking. Initially, these theories placed more emphasis on describing the necessary appraisal criteria than spelling out the algorithm that mapped appraisal input to emotion output [1]. Such specification becomes unavoidable when one wishes to operationalize a given appraisal theory as a computational model. This is what several studies have attempted, with the OCC model of appraisal being the most widely used [3]. The aim of computational models of appraisal (CMAs) is to validate a theoretical model by direct simulation. The advantage of CMAs are that they make explicit assumptions about 1) what is input and output, 2) the data format, and 3) how input maps to output. Typically, CMAs will directly implement the prediction tables of an appraisal theory as a matching algorithm. In that sense, such models adhere to the interactional assumption of appraisal theories. The major disadvantage of CMAs is that they may be overdetermined by theory. One could question...
whether theory alone is capable of fully specifying the relationship between appraisal and emotion. The actual relationship may be more or less complex than assumed. To illustrate this point, let us consider again hierarchical trees as a model for the appraisal-emotion relationship. While many appraisal theories can be formulated in this manner, it is well known in statistics that tree-structured models exhibit high instability [13]. Barring the top-level main effect, a tree essentially models only interactions. If the true appraisal-emotion mapping is, in fact, simple additive, a tree-structured model will have difficulty correctly representing this mapping [13]. From this point of view, a decision tree such as the OCC model could be overly complex.

Directly operationalizing an appraisal theory and simulating results is one approach to validating appraisal theories. Recent work has even attempted to formalize different theories in a common language to facilitate systematic comparison [14], [15]. Broekens and colleagues [14], for example, used this strategy to compare Scherer’s model of appraisal to that of Smith and Kirby [16]. Simulation can tell us when and where the operationalized theory fails but, unfortunately, the method typically offers little more flexibility to detect alternative mappings. An alternative approach to modeling therefore is to collect data and allow statistical methods to estimate a correct model. This has been the dominant approach to validating appraisal models in psychology literature thus far, and will be discussed in the following section.

1.2.2 Data-Driven Modeling

The aim of data-driven methods is to estimate the appraisal-emotion relation bottom-up, rather than imposing a model top-down (as CMAs do). The basic idea is to collect emotion data and let an appropriate statistical model detect the relevant appraisal-emotion associations/mapping. This approach is attractive because it makes fewer assumptions about the nature of the appraisal-emotion relation. There are still assumptions inherent in the choice of variables, which are determined by theory, and in the choice of statistical method. The latter choice has been largely overlooked in the literature thus far. As previously mentioned, an interaction between two variables is a type of nonlinear effect. All major appraisal theories predict such effects for differentiating emotions. To what extent has data-driven modeling respected this assumption? Below we briefly review the literature. This literature can be classified into two domains: 1) research on the overall discriminative strength of appraisal on emotion, and 2) research on specific appraisal determinants of specific emotions.

Many studies have investigated the overall discriminative strength of appraisal on emotion [17], [18], [19], [20], [21], [22]. The cited references all used linear discriminant analysis (LDA) to model the appraisal-emotion relation. Reported accuracy varies between 32 and 80 percent (i.e., the number of cross-validation cases classified correctly). In these studies, the predictor variables were continuous appraisal variables, whereas the response variable was a categorical emotion label (with each category level corresponding to a specific emotion class, e.g., fear, anger, joy). Linear discriminant analysis is a model that attempts to differentiate the response classes by constructing additive linear boundaries in the predictor space (Fig. 3). However, this restriction can be alleviated by including interaction terms between predictor variables. Unfortunately, only one of the aforementioned studies considered interactions between appraisals [20]. In a footnote, the authors mention finding no evidence of improved classification performance with interactions.

By far the most appraisal research has focused on linking specific appraisal determinants to specific emotions. This research can be categorized according to whether appraisals were manipulated (e.g., [23], [24], [25]) or observed (e.g., [26], [20], [27], [28]). Studies that manipulate appraisals typically employ ANOVA interaction designs, where one or more appraisal criteria are varied across levels of other appraisal criteria. The aim is usually to show that specific combinations of appraisals lead to specific emotions, validating the type of prediction tables produced by appraisal theorists. In that sense, such studies are quite similar to CMAs, in that they incorporate interaction assumptions by directly operationalizing theoretical tables. Unfortunately, ANOVA interaction designs have size constraints, in that the number of conditions grows exponentially with the number of appraisal criteria. For example, for six appraisal criteria (an average across theories, [4]), with binary data format (0 = appraisal absent, 1 = appraisal present), a full ANOVA design requires 64 conditions. With ternary format (−1 = appraisal negative, 0 = appraisal absent, 1 = appraisal positive), a full ANOVA design requires 729 conditions.

Studies that use observational data tend to focus on non-interactional analyses, such as additive linear regression. This allows a model to incorporate continuous appraisal variables and also more appraisal criteria than experimental ANOVA designs. The aim of observational studies is usually to investigate one-to-one relations between specific appraisals and specific emotions. The drawback of these studies is the linear nature of the appraisal effects. Interactions are typically not considered and may be difficult to select if the number of predictor variables is large. For example, for six appraisal predictors, there are 15 possible two-way interactions, 20 possible three-way interactions, and so on. Although not unfeasible, classical regression models do not facilitate the detection of such interactions.

The studies just reviewed were mainly concerned with confirming that there is an influence of appraisal on emotion. No formal or systematic attempt was made at comparing different models and mapping algorithms. The choice of model was arguably a matter of tradition (ANOVA) or convenience (LDA), often without explicit acknowledgement of the implied assumptions. It seems clear, however, that a purely additive method cannot be considered a satisfactory model of analysis, given the interactional hypotheses of appraisal theories. Likewise, factorial ANOVA designs impose restrictions on the number of variables to be considered, as well as their data format, favouring two or three appraisal criteria in binary format. If we wish to better grasp the appraisal-emotion relationship, more sophisticated tools seem called for. We now propose to seek these tools in the field of machine learning.
1.2.3 The Machine Learning Solution

In the previous sections, we have argued that traditional approaches to modeling the appraisal-emotion relationship carry restrictions. Current computational models of appraisal conform to theoretical assumptions but lack the flexibility offered by data-driven methods. Data-driven methods offer flexibility but current research has not taken advantage of these possibilities, or even neglected to account for interactional hypotheses of appraisal theory.

The field of statistical machine learning—also known as pattern recognition or data mining—is concerned with extracting complex patterns from large data sets. Frequently, the goal of machine learning is to accurately predict a certain response variable (e.g., emotion class) based on one or more predictor variables (e.g., appraisal criteria). Many models of machine learning exist [13], covering basic methods such as linear regression and tree models, as well as more sophisticated methods such as artificial neural networks or support vector machines. Typically, machine learning does not restrict data analysis to just one model, but rather compares many models, and chooses the one that achieves the best predictive accuracy. The better the predictive accuracy of a model, the more likely the model captures the underlying patterns of the data. A common challenge in data analysis is the presence of non-linear patterns in the data, such as curves and interactions between predictor variables. Thus, if a nonlinear statistical model would achieve a better predictive accuracy on appraisal-emotion data, it would argue against a linear mapping of appraisal to emotions.

What are the advantages of applying machine learning models to appraisal-emotion data? We list the following possibilities:

- Comparing appraisal-emotion models with varying degrees of nonlinearity in a data-driven fashion.
- Including many appraisal criteria without restrictions on the data format.
- Automating the search for meaningful interactions between appraisal criteria, even when the number of appraisal criteria is high.
- Obtaining a complex model without (necessarily) sacrificing interpretability.

Models of machine learning have a reputation for being uninterpretable black boxes. This is certainly true for some (e.g., support vector machines) but not all models. Some methods were, in fact, created with the explicit purpose of obtaining an interpretable model (e.g., trees, multivariate adaptive regression splines (MARS)).

Note that the application of machine learning in emotion research is not new. Studies on automatic emotion recognition, for instance, have used machine learning algorithms to classify emotions based on facial or vocal expression data [29]. By contrast, the application of machine learning to the study of emotion production has not found widespread use. Two relevant studies can be cited.

Tong et al. [30] manipulated the absence or presence of multiple appraisals at different levels and modeled subjectively reported emotion intensity as a function of these appraisals. Although the study was not cast in the machine learning framework, the authors did attempt systematic data-driven modeling by considering regression models with higher-order polynomial terms such as quadratic and cubic terms (e.g., \( y = x + x^2 + x^3 \)). This work was inspired by the theory of Kappas [31], who proposed sigmoidal curves (s-shaped) to model the relation between appraisal and emotion intensity. Tong et al. found evidence of nonlinear relationships between seven appraisals and five emotions across different subject pools, cultures, and sampling methods. Emotion was treated as a continuous response variable (e.g., anger intensity), and linear models were contrasted with nonlinear models. Note that the polynomial terms in this study can be considered interactions of a variable with itself. A significant quadratic effect of a predictor \( x \), for example, means that the relation between \( x \) and \( y \) is modified by levels of \( x \) itself. Tong et al. did not consider interactions between different appraisal variables, however.

Nguwi and Cho [32] used the ISEAR data set (International Survey on Emotion Antecedents and Reactions [33]) to differentiate seven emotions (joy, fear, anger, sadness, disgust, shame, guilt) based on self-reported information of subjective feeling, physiology, action tendencies, motor expression and appraisal (40 items in total). Using a linear support vector classifier, they ranked the predictive strength of each item for each emotion. Results showed that appraisal items were among the highest ranked predictors in the data, with at least one appraisal item in the top three features for each emotion class. The interest of Nguwi and Cho was mainly in deriving a self-organized visual map of high-dimensional emotion data. They did not systematically compare linear to nonlinear methods, nor did they investigate the discriminative strength of appraisals separately from the other emotion components.

1.3 Current Research Objectives

For the present study, we wished to expand the pioneering work of Tong et al. and Nguwi and Cho by conducting a systematic comparison of machine learning models on appraisal-emotion data. To achieve this, we obtained a large data set \((N = 5901)\) from a web-based expert system on recalled emotion episodes [34], with 25 appraisal predictors and a categorical emotion response (12 levels). The goal of the analysis was threefold:

1. Determine whether nonlinear models achieve a better predictive accuracy for the data than linear models.
2. Investigate the overall discriminative strength of appraisal on the 12 emotion categories.
3. Identify specific main effects and interactions of appraisal variables on emotion differentiation.

For appraisal theory, the main issue at stake in these analyses was whether interactions between appraisal criteria really matter. Do we need to account for interactions in our model, as theory predicts? If so, how are the interactions structured and what is their relative contribution to emotion differentiation? If not, is the best mapping nonlinear in some other way, or simply linear additive? A number of methods are available to evaluate the need and contribution of interactions, which will be discussed in detail in the method section. For now, we note that our modeling strategy was largely guided by the interaction hypotheses of...
appraisal theory, which suggest that the appraisal-emotion relation has the following features:

- The relation is characterized by interactions between appraisal criteria.
- Interactions have a hierarchical structure.
- The further down the interaction hierarchy, the more specific emotion differentiation becomes.
- The top-level main effect concerns appraisal of goal-compatibility.
- Lower-level interactions concern appraisal of coping potential, agency, and/or norms.

The method section details how these hypotheses were investigated in the present data.

Which aspects of emotion production did we model in our data? According to the componential perspective on emotion (Fig. 1), appraisal affects changes in motivation, physiology, expression, and feeling. In this study, we focus uniquely on feeling as a response, as represented by discrete emotion labels (e.g., pride, irritation, despair). While this is obviously a reductionist perspective on emotion production [35], the nature of the data set did not allow modeling of additional response variables. Moreover, this approach conforms to past research that also used discrete emotion labels as a response.

2 Method

2.1 Data

Data were obtained from a recent study on appraisal and emotion [34]. In this study, 6,034 users responded to a web-based expert system called the Geneva Emotion Analyst (GEA; see [36], for an earlier application of this system). The GEA first asks respondents to recall “an intense emotion experience that was brought about by an event.” Next, users can label the experienced emotion using maximally two labels—a primary and a secondary choice—from a fixed list of 13 that includes pride, joy, pleasure, rage, irritation, contempt, disgust, shame, guilt, anxiety, fear, sadness, and despair. Finally, respondents are asked to rate the extent of time that caused the emotion on 25 appraisal items (Table 3), corresponding to major categories of appraisal, such as relevance (to what extent is the event relevant), implication (what are the consequences of the event), cause (who or what caused the event), coping (is there power to deal with the consequences of the event), and normative significance (how does the event correspond to social norms and moral standards). Each item is measured on a six-point scale assessing to what extent the appraisal was or was not applicable to the event that caused the emotion, ranging from “1 = not at all” to “3 = moderately” to “5 = extremely,” with “0 = not applicable.” The data set resulting from this study contained 6,034 independent observations on 25 appraisal variables, each case labeled with a primary and a secondary emotion label. Prior to any analysis, 133 cases were removed, either due to missing data (four cases), response bias (61 cases), or low sampling frequency (all 68 cases with “contempt” as primary label), leaving a sample of 5,901 observations. Gender in the data was distributed unequally, with 3,943 females and 1,958 males. The age of the users ranged from 12 to over 60, with the majority aged between 20 and 40 years (approx. 60 percent). Three different language groups were represented in the data, English speaking (2,740), French speaking (2,575) and German speaking (586). Finally, we note that participants in this study were self-selected, as the GEA tool was freely available on the website of the Swiss Center for Affective Sciences, with data being gathered during the course of several years.

2.2 Software

All analyses were carried out with the R statistical software, version 2.14.1. A list of packages that were used to fit each model is provided in the supplemental material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TAFFC.2013.25.

2.3 Data Analysis

2.3.1 General Procedure

The data analysis proceeded in three stages: 1) cluster analysis, 2) black-box modeling, and 3) appraisal feature selection. For the cluster analysis, we first obtained mean appraisal profiles by calculating the mean value of each appraisal variable for each emotion class, yielding 12 vectors of 25 means (i.e., class centroids). The structure of these centroids was then inspected by performing a hierarchical cluster analysis on the euclidean distance matrix. For clustering, we used agglomerative clustering with Ward’s method. The aim of the cluster analysis was to find out to what extent the 12 emotion labels could be grouped into families (e.g., positive and negative emotions).

In the black-box modeling stage, we modeled primary emotion choice (categorical, 12 levels) as a function of the 25 appraisal variables (ordinal, 0-5). The data were first split into a training sample and a validation sample. For the training sample, 100 cases of each emotion were drawn at random from the data set ($N_{\text{train}} = 1,200$). The remaining cases were assigned to the validation set ($N_{\text{val}} = 4,701$). In total, 14 different models were fitted to the training data, four of which were linear and 10 of which were nonlinear (Table 1). The 14 models were selected to cover a wide range of different tools for machine learning, including traditional approaches (linear discriminant analysis, multinomial logistic regression), pure black-box learners (artificial neural networks, support vector machines), hierarchical models (classification tree, multivariate adaptive regression splines), and theoretical prototypes (Scherer prototypes; re-applied from [34]). A detailed discussion on the mathematical properties of these models is beyond the scope of this paper, but an extensive treatment can be found in [13]. For a given model, control parameters were optimized by a five-fold cross-validation on the training data. A full list of the final model specifications is included in the supplemental material (Table S1), available online.

The predictive performance of a fitted model was calculated as the bootstrapped average hit rate across the 12 emotion classes in the validation set. By calculating accuracy only for the validation data, we wished to avoid problems of overfitting. For example, some machine learning models are

---

1. Response profiles were considered biased when only 2 or fewer unique response values were used throughout the questionnaire, or when over 70 percent of responses were of the “not applicable” kind.
In the present study, we extended this approach by first expanding the data set with all possible 300 two-way interactions between the appraisal variables (making 325 predictors in total). We then applied the LASSO criterion to the expanded data set, so as to identify both relevant main effects and interactions in differentiating emotions. This procedure returned 12 regression equations, one for each emotion category, which have to be interpreted in a one-versus-all fashion (see results). Finally, we plotted the obtained structure in a regression diagram for selected emotions.

Note that we avoided classical significance testing in this study for several reasons: 1) the size of the data set ($N = 5,901$) would render almost any association significant due to the high statistical power, 2) conducting 325 separate tests represents a large multiple testing problem, 3) the LASSO criterion is sensitive to effect size rather than statistical significance, and this is what we are ultimately interested in.

### 2.3.2 Testing Interactions

The need and contribution of interactions between appraisal criteria was tested both in the black-box modeling stage and the feature selection stage.

In the black-box modeling stage, models were first ranked by their predictive accuracy. As a general test of nonlinearity, the top nonlinear model was compared to the top linear model via a bootstrapped paired $t$-test. Next, we inspected performance of specific models. Of particular interest were the two hierarchical models: the classification tree, and multivariate adaptive regression splines. Both of these models build a hierarchical model with a high degree of interactivity, although MARS has more flexibility in capturing additive patterns. Since the prediction tables of most appraisal theories can be reformulated (or have been formulated) hierarchically, one would expect the classification tree or MARS to achieve an especially high accuracy. We tested this by inspecting performance of these models, as compared to alternative models.

The hierarchical interaction hypothesis was also tested with an alternative approach. Once the best nonlinear (interactive) model and the best linear model were identified, we repeated the black-box analysis for the emotion clusters uncovered by the hierarchical cluster analysis. According to appraisal theory, interactions should become more important as emotion differentiation becomes more specific. We may not need interaction terms if we only wish to differentiate positive from negative emotions, but need them for very specific differentiation, such as rage versus irritation. In other words, as the number of response clusters to be differentiated approaches the number of specific emotions in the study (12), one would expect the advantage of nonlinear models over linear models to increase.

In the feature selection stage, we first inspected how many interaction terms (if any) were selected as opposed to main effects terms. Next, we quantified the relative contribution of these groups with measures of residual deviance and adjusted McFadden’s $D$, which is a categorical approximation of continuous $R^2$ (a so-called pseudo-$R^2$ measure of explained variance). As in the black-box stage, these analyses were repeated for the emotion clusters uncovered in the hierarchical cluster analysis.

### 3 Results

#### 3.1 Cluster Analysis

We first obtained mean appraisal profiles by calculating the mean value of each appraisal variable for each emotion class, yielding 12 vectors of 25 means. These 12 vectors represented the so-called class centroids. A hierarchical cluster analysis on the euclidean distance matrix of these centroids revealed four main clusters (Fig. 2). The dendrogram in
3.2 Black-Box Modeling

Results of the black-box modeling are summarized in Table 2. As expected, nonlinear models achieved a better predictive accuracy for these data than linear models, with the random forest classifier placed first at an average bootstrapped accuracy of 27.90 percent. This model was followed by the support vector machine (27.18 percent), and LASSO multinomial regression with interactions (26.15 percent). The best linear classifier, LASSO multinomial regression without interactions, followed in fourth position at 25.51 percent accuracy. A paired $t$-test revealed that the random forest was significantly more accurate than both the second best model (SVM), $t(999) = 17.2, p < 0.0001, d = 0.77$, and the best linear model (L-MLR), $t(999) = 65.1, p < 0.0001, d = 2.05$.

Among the linear models, the LASSO multinomial regression model performed the best. Recall that this model is a multinomial regression model with automatic elimination of irrelevant predictors. The comparatively good performance of this model suggested that, for linear models, some appraisal variables were indeed redundant for differentiating these emotions. The Scherer prototypes performed the least well among the linear models and the least well overall at 17.30 percent accuracy (10 percent less accurate than the random forest), although this was still above the chance level of 8.33 percent.

Among the nonlinear models, the two hierarchical models—the classification tree and MARS—performed surprisingly bad. This result argued against a strongly hierarchical appraisal-emotion mapping, as predicted by some appraisal theories (e.g., OCC model). In fact, although the MARS model was allowed to include up to sixth degree interactions, no interaction terms higher than second degree were selected. The classification tree did not perform well, yet the best performing model for these data was the random forest, which uses ensembles of classification trees (fitted on bootstrapped data) to generate predictions [39]. How can we reconcile this result with the poor performance of the single classification tree? And what does the random forest tell us about the appraisal-emotion relation?

In Section 1.2.1., we already noted that classification trees are known for their instability [13]. Due to their highly interactive nature, small perturbations in the data may produce radically different tree-structures between samples. Such deviations are typically more dramatic for increasing interaction-depth (e.g., 4th degree interactions). The random forest counteracts this instability by allowing many trees to vote on the final prediction, simulating perturbations of data via bootstrapping. When presented with new input data, each tree in the forest will generate a prediction for the output. The final prediction of the random forest is then decided by majority vote: if the majority of
the trees in the forest predict a certain class (e.g., fear), then this class is taken as the final prediction. The process of majority voting makes the random forest highly adaptive to nonlinear boundaries between response classes, while filtering out unstable interactions found by a single classification tree.

We can visualize the emotion boundaries estimated by the random forest by constructing a grid of points and plotting predictions for these points in a reduced (principal) appraisal space. Fig. 3 compares the estimated boundaries of the random forest with those constructed by LASSO multinomial regression without interactions (L-MLR). From this figure, we see that L-MLR simply constructed one polygonal area for each emotion class (Fig. 3, right panel). The random forest, on the other hand, had no restrictions on the number or the shape of areas per emotion class. The figure for this model suggested that the boundary between the positive and the negative emotions was virtually perfect, with all models at approximately 93 percent average bootstrapped accuracy. The best nonlinear model was significantly better than the linear model, \( t(999) = 33, p < 0.0001, d = 1.04 \), although the practical difference was negligible. Accuracy declined when we attempted to differentiate the four emotion clusters (happiness, anger, shame/guilt, distress), with all models at approximately 60.5 percent average bootstrapped accuracy. The best nonlinear model was significantly better than the linear model, \( t(999) = 21.9, p < 0.0001, d = 0.69 \), although again, the practical difference was negligible. Accuracy declined further when we attempted to differentiate the

Differentiation between positive and negative emotions was virtually perfect, with all models at approximately 93 percent average bootstrapped accuracy. The best nonlinear model was significantly better than the linear model, \( t(999) = 33, p < 0.0001, d = 1.04 \), although the practical difference was negligible. Accuracy declined when we attempted to differentiate the four emotion clusters (happiness, anger, shame/guilt, distress), with all models at approximately 60.5 percent average bootstrapped accuracy. The best nonlinear model was significantly better than the linear model, \( t(999) = 21.9, p < 0.0001, d = 0.69 \), although again, the practical difference was negligible. Accuracy declined further when we attempted to differentiate the

Table 3 compares performance of the random forest and the two LASSO models (with and without interactions) for varying response clusterings. What can we conclude based on this table? Firstly, that predictive accuracy decreased as the number of emotion clusters to differentiate increased, and secondly, that the contribution of interactions became more sizeable as differentiation became more specific.

![Fig. 3. Estimated decision boundaries of the Random Forest classifier (left) versus the LASSO multinomial regression without interactions (right) in the principal appraisal space.](image)

**TABLE 3**

<table>
<thead>
<tr>
<th>Clustering</th>
<th>Model</th>
<th>Effect</th>
<th>( N )</th>
<th>Deviance</th>
<th>( D )</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 emotions</td>
<td>LASSO</td>
<td>Main</td>
<td>4</td>
<td>4860.59</td>
<td>0.185</td>
<td>25.5%</td>
</tr>
<tr>
<td></td>
<td>LASSO</td>
<td>Interaction</td>
<td>36</td>
<td>4038.40</td>
<td>0.323</td>
<td>26.2%</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>27.9%</td>
</tr>
<tr>
<td>Clusters</td>
<td>LASSO</td>
<td>Main</td>
<td>4</td>
<td>1475.47</td>
<td>0.335</td>
<td>60.4%</td>
</tr>
<tr>
<td></td>
<td>LASSO</td>
<td>Interaction</td>
<td>23</td>
<td>1310.20</td>
<td>0.409</td>
<td>61.0%</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>60.1%</td>
</tr>
<tr>
<td>Pos/Neg</td>
<td>LASSO</td>
<td>Main</td>
<td>3</td>
<td>355.65</td>
<td>0.743</td>
<td>93.0%</td>
</tr>
<tr>
<td></td>
<td>LASSO</td>
<td>Interaction</td>
<td>2</td>
<td>342.97</td>
<td>0.753</td>
<td>93.1%</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>93.2%</td>
</tr>
</tbody>
</table>

Overall results of predictive accuracy for black-box modeling and appraisal feature selection combined, showing the classification accuracy on the test data for the three models considered (Acc.). For the LASSO models, the number of selected main effects and interactions, residual deviance, and adjusted McFadden’s  \( D \) (pseudo- \( R^2 \) measure of explained variance) is shown. RF = Random forest.
12 original classes, but this time nonlinear models gained a more sizeable advantage over linear models. The best nonlinear was significantly better than the linear model, $t(999) = 65.1, p < 0.0001, d = 2.05$.

### 3.3 Appraisal Feature Selection

Table 3 displays overall results of the appraisal feature selection stage. Recall that, in this analysis, we used LASSO multinomial regression to identify specific main effects and (if any) second degree interactions to differentiate specific emotions. As in the black-box modeling stage, we repeated the analysis for varying response clusterings.

The overall results supported the initial findings of the black-box modeling. The LASSO model always selected meaningful interaction terms, even for differentiating positive from negative emotions. As the number of response clusters increased, however, the model selected more interaction terms, rising from 2 to 23 to 36 terms. For differentiating positive from negative emotions, the two interaction terms were relevant but did not contribute much in terms of explained variance, with adjusted McFadden’s $D = 0.743$, for main effects only, and $D = 0.753$, for main effects and interactions. For differentiating the clusters, the interaction terms contributed more explained variance, with $D = 0.335$, for main effects only, and $D = 0.409$, for main effects and interactions. For differentiating all emotions, the difference in explained variance was the largest, with $D = 0.185$, for main effects only, and $D = 0.323$, for main effects and interactions.

The advantage of using interactions in the LASSO model was more pronounced for adjusted McFadden’s $D$ than for classification accuracy (Table 3). Note that these two measures represent a somewhat different approach to assessing predictive success. Adjusted McFadden’s $D$ measures prediction success using only the training set but penalizes for the number of parameters to prevent overfitting, like AIC or BIC. Classification accuracy measures prediction success using only the validation data. Hence, some divergence between these two measures is expected.

The LASSO multinomial regression yielded equations of coefficients, just as in ordinary regression. Before discussing specific equations, however, we first take a look at which appraisal variables were selected overall, at which level (main effect or interaction), and how often (for interactions). Table 4 summarizes these results for each response clustering. When we focused on differentiation between all 12 emotions, two findings were noteworthy. The first finding was that, with the exception of CONSEQUENCES FELT, all appraisal variables appeared at least once in the final model, either as a main effect or as part of an interaction. The second finding was that appraisal variables differed strongly in their degree of interactivity. Intrinsic PLEASANTNESS was the most interactive variable, appearing in 11 interaction terms, in addition to its main effect. Agency-related appraisals were also highly interactive, such as CAUSE-SELF, CAUSE-CHANCE, and degree of OTHER- or SELF-INTENTIONALITY, appearing in eight, five, three, and two

---

**Fig. 4. Regression diagram for predicting positive versus negative emotions.**

TABLE 4

| Clustering | Effect type | RELEVANCE | PLEASANTNESS | UNPLEASANTNESS | OBSTRUCTIVENESS | PREDICTABILITY | CAUSE-CHANCE | CAUSE-SELF | CAUSE-OTHER | INTENTIONALITY | CONS EXPECTED | CONS NEAR FUTURE | SELF-MORPH COMP | MORAL VIOLATION | UNFAIRNESS | ACCEPTABILITY | MODIFIABILITY | ADJUSTMENT |
|------------|-------------|-----------|--------------|----------------|----------------|----------------|--------------|-------------|-------------|---------------|---------------|-----------------|----------------|---------------|-------------|------------|--------------|-------------|-------------|
| All emotions | Main | 1 | 1 | 1 | 1 | | | | | | | | | | | | | | |
| | Interaction | 2 | 11 | 4 | 2 | 3 | 2 | 1 | 3 | 5 | 8 | 2 | 1 | 3 | 1 | 1 | 1 | 3 | 4 | 4 | 3 | 1 | 1 |
| Clusters | Main | 1 | 1 | 1 | | | | | | | | | | | | | | | | | | | | | | | | |
| | Interaction | 5 | 2 | 4 | 2 | 2 | 1 | 2 | 1 | 2 | 5 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 3 | 2 | 4 | 1 | | | |
| Pos/Neg | Main | 1 | 1 | 1 | | | | | | | | | | | | | | | | | | | | | | | | |
| | Interaction | 1 | 1 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | |

Selection of appraisal variables at the level of main effects and/or interactions by response clustering, as determined by LASSO multinomial regression. Numbers refer to the amount of interactions that this variable appears in. For main effects, the default is always 1. Empty cells indicate that this variable was not selected for the corresponding effect. Cons. = consequences, comp. = compatibility.
interaction terms, respectively. Norm-related appraisals were moderately interactive, such as NORM VIOLATION, UNFAIRNESS, and MORAL ACCEPTABILITY, appearing in four, four, and three interaction terms, respectively.

Let us now examine specific regression equations found by the LASSO model. In the following paragraphs, we make use of regression diagrams to represent selected models. This we did to enhance interpretability and to explicitly represent the LASSO equations in a computational format. Such models could easily be exported for use in other applications. Because there are 17 potential diagrams to depict (pos/neg, four clusters, 12 emotions), we focused on interpreting five specific diagrams instead: positive versus negative emotions, the shame/guilt cluster, and the emotions of rage, fear, and despair, specifically. For the full coefficient table of the LASSO model, we refer the reader to the supplemental material, available online.

Fig. 4 depicts the regression diagram for differentiating positive from negative emotions. The response in this model was the log-odds of a positive emotion occurring versus a negative emotion occurring. Solid arrows on the diagram indicate positive associations, while dotted arrows indicate negative associations. Main effects are depicted as straight arrows influencing the response variable. From Fig. 4, for example, we see that there was a strong main effect of intrinsic PLEASANTNESS on the probability of a positive emotion occurring. More specifically, when all other appraisals were set to 0, an increase of appraised intrinsic pleasantness by 1 unit increased the expected log-odds of a positive emotion occurring by 0.946. This corresponded to a multiplication in raw odds by 2.575. Similar but reverse main effects were found for intrinsic UNPLEASANTNESS and OBSTRUCTIVENESS. Interactions in the regression diagram are represented as round-headed arrows modifying other arrows. Some of these interactions, such as PREDICTABILITY × OBSTRUCTIVENESS, modified an existing main effect (OBSTRUCTIVENESS). In this case, increasing appraisal of predictability of an event reinforced the negative effect of goal obstruction appraisal. Others interactions were pure interactions lacking a main effect, such as CONDUCTIVENESS × OTHER-INTENTIONALITY. A pure interaction meant that, for non-zero values, the two variables jointly influenced the probability of a positive emotions occurring. One variable could not influence emotion if the other had a zero-value. The intercept, finally, represented the log-odds of a positive emotion when all appraisal variables in the model were set to 0. A log-odds of −1.382 corresponded to odds of 0.25 of a positive emotion occurring versus a negative emotion occurring, or a probability of 20 percent. In other words, when no appraisal was active, a negative emotion was therefore the model's best guess.

Fig. 5 (upper left) depicts the regression diagram for the shame/guilt cluster. For this model, the response was the log-odds of shame/guilt occurring versus an emotion from the happiness, anger, or distress cluster. The diagram highlights the importance of own agency for predicting shame and guilt, as represented by CAUSE-Self and INTENTIONALITY. For self-causation, the LASSO model selected a main effect
and three interactions: cause-self × norm violation, cause-self × avoidability, and cause-self × modifiability. All these interactions reinforced the main effect of self-causation. Rather than interpreting the coefficients directly this time, we can also enter values for the appraisal variables and directly compute the model’s predicted probability of shame/guilt occurring. When cause-self, norm violation, avoidability, and power were all at their minimal value of 0, the probability of shame/guilt occurring versus the three other emotion clusters was approximately 39 percent. When they were all at their maximal value of 5, the probability of shame/guilt occurring versus the three other emotion clusters was approximately 75 percent. This would happen when an event was appraised to be completely self-caused, highly norm violating, and one that could have been avoided and modified.

Finally, we look at diagrams for the emotions of rage, fear, and despair specifically. Fig. 5 (upper right) depicts the regression diagram for rage. The first thing to note is that rage was entirely determined by interactions (no main effects). The second is that all coefficients were rather small. The largest effect was found for the interaction between other-intentionality and unfairness. For example, when an event was appraised to be highly unfair and caused intentionally by someone else (both values at 5), the odds of rage occurring versus one of the other 11 emotions was 2.33, which corresponded to a probability of 70 percent. The role of norms in eliciting rage was further highlighted by the interaction between norm violation and urgency. The effect was rather small when added to the other-intentionality × unfairness interaction, however, increasing the probability of rage by just 1 percent, even at maximal values of 5.

Fig. 5 (lower left) depicts the regression diagram for fear. As with rage, fear was entirely determined by interactions (no main effects). The most important predictor of fear was cause-chance, an effect that required the co-occurrence of obstructiveness, intrinsic unpleasantness, and urgency appraisals. Urgency appraisal in turn interacted with the appraised suddenness of an event. In plain terms, fear was expected when a highly negative event occurred suddenly and required immediate action. At the maximal value of 5 for these appraisal variables, the odds of fear occurring versus one of the other 11 emotions was 3.62, corresponding to a probability of 78 percent.

Fig. 5 (lower right) depicts the regression diagram for despair. As with rage and fear, despair was entirely determined by interactions (no main effects). This time appraised personal relevance exerted the most influence, although this effect was heavily co-dependent on the three other appraisal variables in the model, cause-self, obstructiveness, and unfairness. When all these variables were set to their maximal value of 5, the odds of fear occurring versus one of the other 11 emotions was 2.52, corresponding to a probability of 72 percent.

4 DISCUSSION

In this study, we conducted a data-driven analysis of the appraisal-emotion relation using models from the field of machine learning. This strategy allowed us to systematically compare different appraisal-emotion mappings without the theoretical restrictions typically faced by computational models of appraisal, and without the practical restrictions faced by traditional analysis methods such as ANOVA and linear regression. Our modeling strategy was guided by interactional hypotheses of appraisal theory, which state that the effect of one appraisal variable on emotion can be modified by another appraisal variable.

Our results support the interactive nature of the appraisal-emotion relation. This finding was evident both in the black-box modeling stage and the appraisal feature selection stage. In the black-box modeling stage, the best linear model was significantly outperformed by the best nonlinear model, and this at three different response clusterings. The best nonlinear model for differentiating all 12 emotions was the random forest, which utilized ensembles of interactive classification trees to generate predictions. In the appraisal feature selection stage, we used LASSO multinomial regression to automatically detect relevant second degree interaction terms for differentiating 12 emotions. For each response clustering, meaningful interactions were selected in the final model. For differentiating the 12 emotions, the LASSO criterion retained four main effects and 36 interactions. Main effects involved appraisals of intrinsic pleasantness and obstructiveness. Interactions largely involved appraisals of agency and norms (Table 4). Simply put, who causes an event and whether or not norms are violated strongly modified the effect of other appraisals on emotion. We now discuss these findings in-depth with respect to three areas: 1) interactions in appraisal theory, 2) emotion differentiation, and 3) modeling emotion production.

Interactions in appraisal theory. As mentioned, our results supported the interactive nature of the appraisal-emotion relationship. Some nuances have to be added. For a start, interactions became less important as emotion differentiation became less specific. The LASSO model did select interaction terms for differentiating positive from negative emotions, but the practical value of these terms with respect to predictive accuracy or explained variance was negligible (Table 3). Interactions mattered when we wished to differentiate very specific emotions, such as guilt, fear, or despair. Second, the degree of interactivity found in these data was not as high as predicted by theory. We did not find evidence for strongly hierarchical interaction structures, which have been proposed (or implied) by appraisal theories (e.g., Roseman, OCC). Among the fitted models, the classification tree and MARS ranked near the bottom in terms of predictive accuracy. Moreover, the MARs model did not select interaction terms beyond the second degree. On the other hand, evidence for strong additivity was also rejected by our data. In the appraisal feature selection stage, only a handful of main effects were selected by the LASSO model. The majority of selected effects were interactions. In that sense, one could characterize the appraisal-emotion relationship for these data as moderately hierarchical. A main separation between positive and negative emotions was first made via intrinsic (un)pleasantness, goal obstructiveness, and relevance. Afterwards, differentiation between specific emotions required second degree interactions. These interactions mainly involved agency and norm related appraisals, as predicted by appraisal theory [4].
In these data, appraisal of goal compatibility was represented by the conductiveness and obstructiveness variables. Most appraisal theories consider these variables fundamental for differentiating positive from negative emotions [4], [7]. In our analysis, models tended to prefer the valence variables as the main effect (pleasantness and unpleasantness), although it should be noted that goal obstructiveness was a main effect in the LASSO regression model, and that the goal compatibility variables correlated strongly with the valence variables (what is goal compatible is often pleasant). Moreover, most appraisal theories still consider intrinsic valence a basic appraisal effect, on the same level as goal compatibility [4]. We did not find a main effect for agency, as predicted by the OCC model [12]. Nevertheless, the importance of agency was highlighted by its strongly interactive nature, with cause-self appearing in eight out of 36 interaction terms in the LASSO model (Table 4). Most appraisal theories also emphasize the role of coping potential as a modifying variable, especially the theory of Lazarus [6]. Our data did not support strong interaction effects of power-related variables. Avoidability and modifiability played a role in differentiating the anger and shame/guilt clusters at the cluster level, but the coefficients tended to be small.

Emotion differentiation. The average bootstrapped accuracy of the best model, the random forest, was estimated to be 27.9 percent, with the maximal bootstrapped accuracy of this model around 31 percent. This percentage exceeds chance level but remains considerably below accuracies reported by other studies, which usually found 40 to 80 percent predictive accuracy. We attribute this discrepancy to the high degree of noise in the GEA data. Firstly, the study was not conducted in a laboratory setting but via the internet, allowing little control over the conditions under which the questionnaire was completed. Secondly, although we took some steps to remove cases with biased response profiles, not all bias could be removed. A repeated measurements protocol would have been required to estimate each respondent’s individual bias. Thirdly, the GEA questionnaire allowed participants the opportunity to provide not just one but up to two emotion labels to characterize the emotion they experienced, which more than 70 percent of respondents did [34]. This made the final data more heterogeneous than the primary label alone would suggest. Ideally, we would have modeled both labels simultaneously, but from a machine learning perspective this task is much more complicated than modeling only one label. Literature in machine learning has generally not focused on models for multiple-response multiclass prediction.

Note also that, in the original Scherer study [36] as well as the recent paper on the theoretical prediction model of the GEA system [34], the theoretical model was allowed to make two guesses rather than one, and calculated an unbalanced estimate of accuracy. At two guesses, the theoretical model is about as good as cluster prediction in the current study (i.e., predicting happiness, anger, shame/guilt, or distress). Typically, the theoretical model can guess the emotion family right, but misses on the specific label. An unbalanced estimate of accuracy gives more weight to more frequent emotion classes, which may be considered realistic. Thus, models that are good at predicting high-frequency classes (e.g., sadness, joy) will obtain a comparatively high predictive accuracy. In this study, on the other hand, we gave equal weight to all emotion classes when calculating final accuracy. This strategy was motivated to maintain the comparability of accuracy scores between different models. Moreover, since the distribution of emotion classes was balanced in the training data, it made sense to calculate a balanced accuracy score for the validation data as well.

The low predictive accuracy in the present study is further nuanced when we consider differentiation between response clusters: positive versus negative emotions, or the four emotion family clusters of happiness, anger, shame/guilt, and distress. Differentiating response clusterings raised predictive accuracy to between 60 and 93 percent, which is much higher than 27.9 percent.

The clusters that were identified in this study corresponded to clusters identified in previous research and theoretical propositions on emotion families. We replicated a happiness and an anger cluster [40], [17], [41], [42]. We did not find evidence for a fear/anxiety cluster that was separate from a sadness/despair cluster. On the other hand, we did find a separate shame/guilt cluster, emotions that are typically clustered with anxiety or sadness. Conceptually, however, these emotion families are highly interpretable. When we take a look at the confusion matrix (Table S3) of the random forest, we also notice that confusions typically took place within clusters, rather than between clusters.

Modeling emotion production. In this study, we successfully applied methods of machine learning to appraisal-emotion data. In doing so, we conducted a systematic data-driven comparison of different appraisal-emotion mappings. We believe that this approach has important advantages over heavily theory-driven strategies applied by computational models of appraisal. While CMAs prove useful for validating a given appraisal theory, they offer little flexibility for detecting alternative appraisal-emotion mappings. As a case in point, consider again the performance of the theoretical Scherer prototypes against the data-driven models for our data (Table 2). Although the theoretical prototypes perform above chance level, the best data-driven model outperforms it by more than 10 percent predictive accuracy. Clearly, there are limits as to what purely theoretical appraisal-emotion models can achieve. On the other hand, we would not advocate a purely data-driven approach to emotion modeling. Science thrives precisely on the interplay between theoretical development and empirical research. Emotion theory plays an important role in defining the object of study, specifying which variables should be included in a computational or statistical model. Estimating relations between these variables, especially if complex relations are proposed, may be best suited to statistical models. Once the data provide new information, we can update the theory anew, and so on. Our results show that it is possible to develop an interpretable model of appraisal that is based on theoretical assumptions while still allowing the flexibility of data-driven methods (see
regression diagrams in Figures 4-5). Unfortunately, research on emotion production so far has rarely taken advantage of machine learning methods. We hope this study will encourage researchers to consider more sophisticated models than simple linear regression or classical ANOVA designs.

Limitations and future research. A drawback of the current study is its rather limited operationalization of emotion. We modeled the relation between appraisal and feeling, treating emotions as static phenomena, summarized “memories” rather than time-varying processes. Recall from Fig. 1 that, according to the componential perspective, emotions are defined as time-varying changes in appraisal, motivation, physiology, expression, and feeling. The current work should be expanded to include more response variables from the emotion domain (motivation, physiology, and expression) and to consider the time-varying nature of emotion. Integrating all these features into a single model remains an important challenge. Again, data-driven methods may be employed to estimate the time-varying relations between the different components of emotion. This requires the application of machine learning models for time series data, which we believe will become increasingly important for correctly modeling emotion processes. We hope that future research continues on the course that we have taken in the current study, which is to consider more sophisticated data-driven approaches to modeling appraisal and emotion.

ACKNOWLEDGMENT

The authors wish to thank Olivier Renaud for insightful discussions regarding the data analysis, Guillaume Chanell for comments on an early draft of the manuscript, and the editors and reviewers for their extensive feedback on the first submission. This research was supported by the National Center of Competence in Research (NCCR) Affective sciences financed by the Swiss National Science Foundation (n° 51NF40-104897) and hosted by the University of Geneva.

REFERENCES

Ben Meuleman received the master's degrees in experimental psychology in 2009 and in statistical data analysis in 2012 from the University of Ghent. He is currently working toward the PhD degree at the Swiss Center for Affective Sciences on a project concerning the computational modeling of appraisal theory of emotion, supervised by Klaus Scherer. The goal of this project is to advance the understanding of emotion production using methods from statistical machine learning.

Klaus R. Scherer holds the chair for emotion psychology at the University of Geneva and directs the Geneva Emotion Research Group. His major research interest is the further theoretical development and empirical test of his component process model of emotion (CPM), specifically the modeling of appraisal driven processes of motor expression and physiological reaction patterns, as well as the reflection of these processes in subjective experience. Other major research foci are the study of the expression of affect in voice and speech, and applied emotion research.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.