Behavioural cues help predict impact of advertising on future sales

Javier Orozco[a], István Petrás[a], Gábor Szirtes[a,*], Dániel Szolgay[a], Ákos Utasi[a], Jeffrey F. Cohn[b]

[a] *Realeyes OÜ, Tölgyfa utca 24, Budapest 1027, Hungary*
[b] *University of Pittsburgh, 4322 Sennott Square, Pittsburgh, PA, USA 15260*

Abstract

Advertising aims to influence consumer preferences, appraisals, action tendencies, and behaviour in order to increase sales. These are all components of emotion. In the past, they have been measured through self-report or panel discussions. While informative, these approaches are difficult to scale to large numbers of consumers, fail to capture moment-to-moment changes in appraisals that may be predictive of sales, and depend on verbal mediation. We used webcam technology to sample nonverbal responses to television commercials from four product categories in six different countries. For each participant, head pose, head motion, and smiling were automatically measured at each video frame and aggregated across subjects. Dynamic features from the aggregated series were input to simple linear ensemble classifier with 10-fold cross-validation to predict product sales. Sales were predicted with ROC AUC= 0.737, 95% CI [0.712,0.762] and were highly consistent among product categories. ROC AUC for product categories varied by less than 1%. Variation was found among countries. Highest results were found for US and France (ROC AUC = 0.82 and 0.86, respectively) and lowest for Australia (ROC AUC = 0.69). In comparison with previously reported results using static features, models using dynamic features yielded higher performance and greater consistency among product categories. These findings support the feasibility, efficiency, and predictive validity of sales predictions from large-scale sampling of viewers' moment-to-moment responses to commercial media.

Keywords: market research, behavioural cue, predictive modelling, facial expression analysis
2016 MSC: 00-01, 99-00

*Corresponding author
Email address: gabor.szirtes@realeyesit.com (Gábor Szirtes)

1. Introduction

Advertising is about influencing consumer preferences, appraisals, action tendencies, and purchases. Television and increasingly online video commercials are a key component. Over 80 billion dollars is spent annually on television commercials in the US alone [1]. For the companies that produce commercials and for their clients, there is great interest in evaluating the effectiveness of commercials they produce and distribute. One approach is to correlate television advertisements with product sales (online shopping in a short time window around the time of tv ad)[2]. This approach enables a gross estimate of direct influence of advertising on sales but is blind to consumer reactions to individual commercials. For that, it is necessary to assess consumer responses to specific commercials in relation to product sales.

One solution is to ask viewers to report on their responses to commercials. Focus groups, personal interviews, random-digit phone surveys, and online surveys have been used for this purpose. While providing useful information, these methods have notable limitations. They pull for rational thinking rather than emotional responses that may be more predictive of purchase behavior; respondents must verbally represent what often are nonverbal, often unconscious cognitive-emotional reactions; and the dynamics of their responses may be compromised by recency effects. Demand characteristics and social desirability effects may bias reports as well. Focus groups, surveys, and related methods further assume that verbal reports are necessarily the best indices of purchasing influences. Evidence suggests otherwise. People's preferences often are outside of their awareness and strongly influenced by emotion [3, 4].

Emotions consist of multiple components that include subjective feelings, action tendencies and physiological arousal. All are prime candidates for influencing likelihood of purchase decisions. During emotion episodes, these components become correlated [5]. Recent work seeks to measure one or more of these indices. Measures of peripheral (e.g., heart rate) central (e.g., EEG and fMRI) physiology and manual annotation of facial expression ([6]) have been explored. These approaches have been useful but are challenging to scale to large numbers of subjects. Electrophysiology requires special equipment, and attaching sensors may inhibit or alter people's responses. Manual annotation of facial expression is labor intensive. None of these methods scale well to the large numbers of respondents needed for population-based estimates. Methods that can be applied to large, representative samples of respondents are needed.

Automated facial expression analysis using webcam video acquisition is a promising alternative. Using computer vision and machine learning, facial expressions of emotion to television advertisements can be measured on a moment-to-moment basis. This approach avoids the necessity for viewers to verbally report their experience, captures fine-grained information about the timing of behavior, and can be scaled to large numbers of viewers from multiple geographical regions and countries. In seminal work, [7] found that facial expression measured in this way was predictive of sales. Given the wide availability of webcam technology and the efficiency of this approach, it becomes possible

to plan population-based research for more accurate investigation of viewer's reactions to commercial presentations and their relation to product sales.

Using sales data from MARS, Incorporated and webcam recordings of viewer responses to commercials, McDuff [8, 9] obtained mixed results. In [9] facial expression based analysis outperformed survey based methods in predicting sales performance, *when average commericals (half of the data) were discarded*. The combined method did not bring about any improvement. In [8] he was able to predict sales performance for three of four product categories evaluated. Accuracy for the fourth was markedly lower than the others. Interestingly, facial expression based method was on par with survey (both achieved moderate accuracy),but their combination achieved significantly higher accuracy. In spite of the mixed results, this work suggested the efficacy of combining "crowdsourcing" methodology, automated facial expression analysis, and supervised machine learning algorithms to differentiate between high and low performing ads where labels are based on sales lift data (i.e., increase in sales). These initial findings suggested that automated behavioral cue based analysis has great practical value and with further improvements this approach can become a viable alternative to traditional, survey based methods. In comparison with alternative methods, it scales well to large samples of respondents and can be executed more quickly. A critical challenge is to achieve higher overall accuracy and greater consistency in performance across product categories.

McDuff's approach emphasizes summary, or static, indices of emotion expression, especially the proportion of smiling, but ignores dynamics of facial expression and their nonverbal context. We wondered whether dynamic features and nonverbal context would result in higher and more consistent accuracy across product categories. In a series of studies, the "packaging" of nonverbal behavior (e.g., co-occurring head pose and motion) and dynamics have proven critical to the meaning of facial expression. Orientation or timing of head movement, for instance, encodes meaning. Smiles of enjoyment and embarrassment, for example, have similar static features (e.g., contraction of the zygomatic major and orbicularis oculi), but differ in head pose and movement. For enjoyment, head pose is frontal or slightly raised, while for embarrassment it pitches down and to the side [10, 11].

Recent work in automatic facial expression analysis suggests that dynamic features (e.g., velocity or acceleration of facial expression and head pose) strongly encode emotion. Velocity and acceleration of head and facial movement, in particular, can express emotion and related affective states [12, 13]. Motivated by these findings, we test the hypothesis that the dynamics of facial expression and head motion in response to commercials is predictive of product sales. We evaluate this hypothesis for the product categories and countries examined by McDuff and include additional countries. We compare our findings for product categories and countries with the previous approach that used only static features to predict sales lift (increase in sales).

The paper is organized as follows. In Section 2 we describe the task and sales data. We then describe the data acquisition method, data representation, and the classification model that uses dynamic features. For each part, differ-

Figure 1: Demographics of the participants (a) age distribution, (b) gender distribution

ences between the dynamic approach and the static one [8, 9] are also given. Performance of the proposed model is then presented in Section 3 together with numerical comparisons with the static approach. In Section 4, we discuss findings and future work informed by those findings.

2. Proposed Method

2.1. Objectives

We recruited population-based samples of viewers in six countries. Viewers were recorded via webcam while they watched commercials on a computer monitor or screen. Commercials were from four product categories. Sales lift by product category and country were obtained from MARS, Incorporated. We constructed predictive models using dynamic measures of facial expression and head motion. Results were compared to ones reported previously that used static measures for the same product categories in the countries included in both studies.

2.2. Participants

Ads were watched online by paid participants selected in a way that samples (panel) demographics follow census statistics of the corresponding country. In addition to demographics constraints, there were 2 more selection criteria. The technical requirement was that each participant has internet access and webcam attached to her home computer. The relevance requirement was that commercials should be displayed to category users only, thus making the ads relevant. This is in contrast to the static approach where only 76% of the participants were actual category users. The total number of participants was 18793, but for quality reasons described in Subsection 2.4 only 12262 sessions were finally used in the analysis. The age and gender distribution of participants is shown in Figure 1 with small differences between countries.

2.3. Stimuli and class membership

The commercials represented four product categories: confections, food, pet care, and chewing gum. They were originally aired between 2013-2015 in the

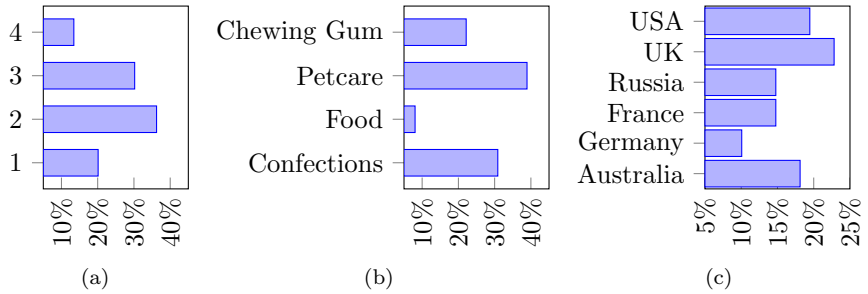

Figure 2: Statistics of the commercials: (a) Distribution of sales-lift ratings from 1 (lowest) to 4 (highest); (b) distribution of product categories; and (c) distribution of countries.

six different countries. The commercials varied in duration between 10 and 30 seconds.

Sales lift data was provided by MARS, Incorporated. Target score was derived from the actual contribution of the ad campaign to "sales lift". To measure sales lift for each commercial, exposed and control (unexposed) comparison groups were identified by MARS, Incorporated and their actual purchases were traced.The ratio of purchase propensity in the exposed group to the comparison group was then averaged over the set of exposed/comparison groups. Sales lift rating was quantified on a four-point ordinal scale for training classifiers. Because distinctions between intermediate levels (2 and 3 on 4-point scale) are problematic, the scale was reduced to a binary representation for classification. The distribution of commercials across categories and regions and the score distributions are plotted in Figure 2.

Complicating analysis, about a third of the commercials were variations of each other. We considered two commercials as variations if differences between them were due to small edits in length or content. As an example, some commercials had the same storyline, but displayed a different brand label or were produced in a different language. We report results separately for all commercials and for the case in which related ads are combined into a single label.

But for a few exceptions, the study design was comparable to [8]. We included two additional countries. The commercials they used aired in 2002-2012; ours aired more recently. Their set contained 163 unique commercials; ours contained 116 unique ones out of the available 149 commercials. Sales lift in their study was quantified on a 3-point ordinal scale and ours on a 4-point ordinal scale.

2.4. Collection of behavioural responses

All commercials were viewed by participants on their own computer while their face was recorded by webcam and streamed to a server. Image resolution was 640×480. This "in the wild" setting ensures more ecologically valid spontaneous behaviour than would be possible in a laboratory at the cost of image quality and frame rate. Average frame rate was about 13 fps. Videos were

Figure 3: Example data

omitted if face was occluded or subjects were engaged in unrelated activities like talking or eating. Figure 3 displays some examples of varying illumination, pose, distance from the webcam, and facial expression.

Subjects viewed up to four commercials presented in a random order. Session length was approximately 10 minutes. By contrast, in [8], subjects watched 10 commercials presented in a random sequence and completed self-report ratings between them; session length averaged 36 minutes. We chose a shorter format because [14, 15] found a negative correlation between session length and data quality. In addition, we used larger samples (on average 284 subjects viewed each ad versus 100) to counter the impact of video quality as well as large variations in the observability of the viewers' responses. Even after applying our conservative quality filtering, the remaining sample size was 161, still significantly larger than 100 as reported for static approach.

2.5. Data representation

2.5.1. Preprocessing

First, color frames were converted into grayscale intensities. Second, facial features were extracted and input to classifiers for emotion detection. Third, the raw features as well as the output of the emotion algorithms were used to form time series signals for predictive modelling.

2.5.2. Facial feature extraction

On each frame we detect the location and pose (yaw, pitch and roll) of the head/face using an improved version of the method of [16] and locate the precise position of a set of facial landmarks (alignment of key points) based on a modified reimplementation of the algorithm proposed in [17]. Local geometry of the landmarks as well as texture patches around them are then used as descriptors by our in-house emotion classification system trained to classify facial expressions into discrete expression categories such as smile, surprise or disgust.

The output is then a multi-dimensional time series of the group memberships as well as the corresponding probability output (posterior probability that a class label is chosen for a given set of descriptors).

Preliminary studies indicated that head pose variation at the *individual level* (signal will be referred to as *Head Pose*) as well as smile probabilities at the *sample level* (signal will be referred to as *Smile Dynamics*) show high correlation (linear Pearson-correlation) with the class labels. In turn we used these signals as input for the predictive model.

As facial features were extracted independently from each frame, we applied zero phase (backward-forward) exponential smoothing to counter noise due to shape misalignment. In addition, to support temporal alignment of the individual responses to a given stimulus, time series were resampled to have uniform frame rate of 10 or 25 fps depending on the signals. Numeric parameters (such as the optimal length of the segment, resampling frequency, filtering parameters) were set to optimize correlation with scores via grid search.

2.5.3. Signal I: Head pose

Because of the large variability of video quality and frame rate, we sought representations of head pose changes that are simple yet informative enough to support predictive modeling. In preliminary testing, we found that head pose variations toward the end of each video had largest correlation with sales lift. We thus chose the sample mean of the variance of individual head poses in the final segment of the recordings. Our assumption is that stronger elicited emotional responses are accompanied with more intensive head pose changes due to head movements. Previous work has found that gaze direction strongly correlates with head pose [18, 19] so larger head pose variations may reflect a lasting effect of the stimulus content and do not correspond to the very last segment of the stimulus, since subjects with extreme head pose do not look at the direction of the screen.

Head pose (yaw, pitch and roll in degrees) was calculated from the estimated 3D shape of the face and the obtained time series were then processed as described above. For each angle the resulting filtered and resampled time series is $h_j \in \mathbb{R}^{3\times k_j}$, $j = 1...n$ where n is the number of recorded sessions belonging to the given advertisement and k_j is the number of measurements in the recording.

A 3 dimensional feature was extracted indicating the expected value of the head pose (yaw, pitch and roll) change in the last segment of the advertisement.

The following signal processing steps are executed:

1. For each session $j = 1, 2, ...n$ let h_j^{end} denote the last m frames of h_j an $\bar{h}_j^{end}$ denote the mean and calculate the standard deviation of the segment: $\sigma_j = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(h_j^{end}(i) - \bar{h}_j^{end})^2}$
2. The extracted feature will then be the standard deviation of the segments averaged over the sessions: $x_h = \frac{1}{n}\sum_{r=1}^{n}\sigma_r$, where $\mathrm{x_h} \in \mathbb{R}^3$

2.5.4. Signal II: Smile dynamics

The most frequent facial expression is the smile [20]. Smiles may convey enjoyment, favorable appraisal, anticipation, and action tendencies to approach. [21]. From perspective of automated detection, smiles often involve relatively large geometric and textural deformations that are advantageous (see e.g. [22]). Since most of the advertisements in our data set were designed to be amusing or joyful, it is expected that signals derived from smile carry information about the elicited emotional states.

Here again the main issue was to find a signal representation that is less sensitive to variation within the sample, yet informative enough to be useful for prediction. While previous attempts relied on simple summary statistics of aggregate sample responses, such as maximum or gradient of a linear fit, we hypothesized that dynamics of the elicited emotional peaks (e.g., the relation between onset, apex and offset) would be more relevant. In turn, our signal is designed to differentiate between emotional events with fast ramp-up and slow decay and events with slow rise and fast decay.

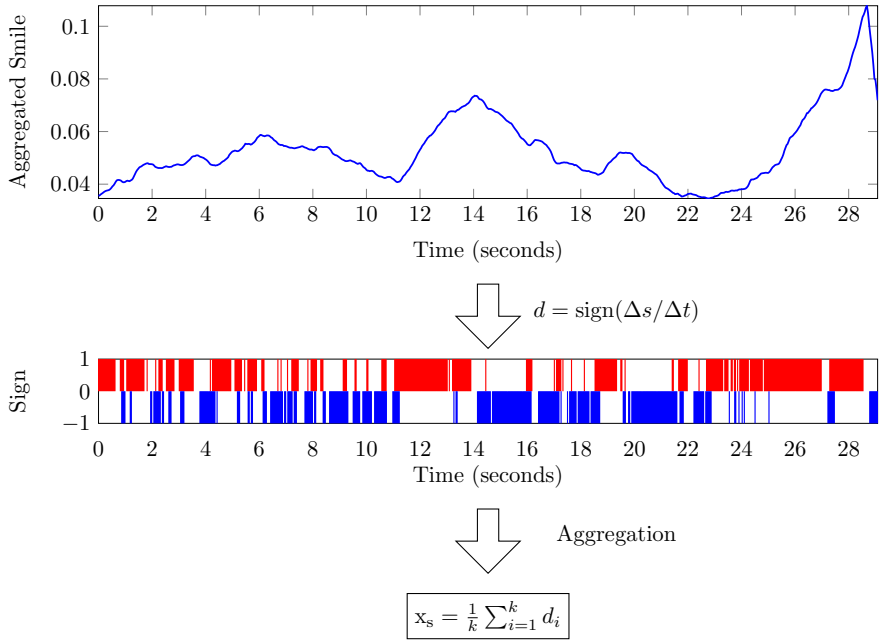

Figure 4: Example of "smile dynamics" signal calculation. First the temporal difference of the aggregate smile response to a given ad is taken (red denotes increase, while blue denotes decrease in the signal) then the sign of the differences is added up to yield one number corresponding to the ad.

We calculate a one dimensional feature s_j from smile probability change along the advertisement:

$s_j \in \mathbb{R}^{1\times k_j}$, $j = 1...n$ where n is the number of recorded sessions belonging to the given advertisement and $k_j$is the number of measurements. Our smile

dynamics signal is generated by the following algorithm:

1. Aggregate the individual smile probability time series into one curve (s), which describes the average smile probability as a function of time: $s\left(l\right) = \frac{1}{n}\sum_{j=1}^{n}\left(s_j\left(l\right)\right)$
2. The signal is then the sum of the sign of the temporal differences, normalized by the number of sample points (duration of the advertisement): $\mathrm{x_s} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{sign}(\Delta s_i/\Delta t)$, where $\mathrm{x_s} \in \mathbb{R}$

An example of smile dynamics signal calculation is shown in Fig 4.

In comparison, in the static approach static signals were extracted from a mix of facial action unit activations which are strongly related to particular discrete expressions (eye brow raise is often associated with surprise), discrete expressions (smile) as well as "valence" which was derived from the estimated intensity of all discrete facial expressions. We instead used a simpler mix of 2 signals related to smile and head pose dynamics.

In the static approach first order summary statistics and gradient of the linear fit were used to represent data with the underlying assumption that rise is an essential component. We argue that slower and somewhat delayed head pose changes may reflect attention as well as the lasting effect of previously seen stimuli so we focused on the last segment of the recordings. Regarding emotional response we also found that smile is the most reliable and most frequent response, however we wanted to capture the dynamics of the emotion response curves. In turn, we did not seek simple positive slope, but analysed the shape of the apex in the panel responses. Referring to the two approaches as Static vs Dynamic is justified by the fact that our data representation better captures dynamic changes in the observable behavioural responses.

2.6. Modelling

Limited sample size and potential label noise makes modelling difficult or even impossible if complexity of the used approach is high. So we opted for simple ensemble modelling with weighted averaging [23, 24] with the following assumptions. We treat signals as independent and do not consider higher order interactions between them. This assumption allows for training simple (weak) experts whose vote can be summarized in an ensemble model. The second assumption is that we seek linear relationships between signals and target score and non-linearity is induced by thresholding (binarization of the individual experts' output). Such thresholding supports signal denoising. The workflow of our model is shown in Figure 5. The ensemble model is composed of standard linear regressors, nonlinear terms (binarization) and a standard logistic regressor to produce calibrated probability outputs. The variables and parameters in the model are indexed with 's' suffix for smile dynamics signal and with 'h' suffix for head pose signal. The signal processing path are the same for both signals, therefore we only describe it for smile dynamics.

The input of the linear regression step is the aggregated Smile Dynamics signal x_s described in Section 2.5.4 (for Head Pose see Section 2.5.3). The target

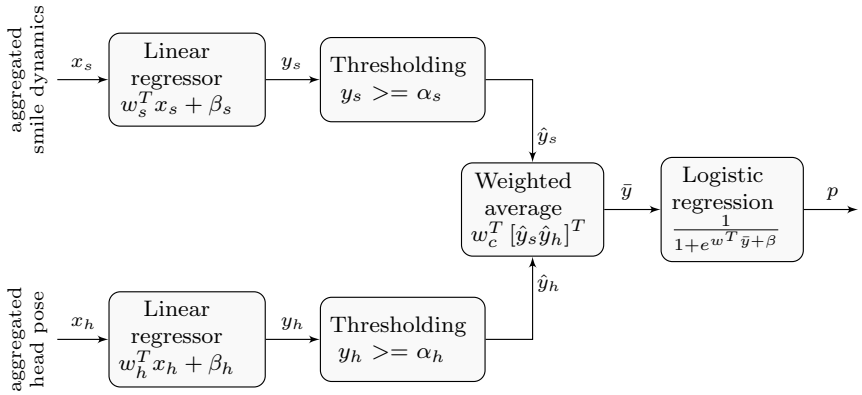

Figure 5: Ensemble predictor. Inputs of the modeling are the aggregated head pose and smile dynamics x_s x_h signals (Section 2.5.4 and 2.5.3). The two branches corresponds to the "smile dynamics" and "headpose" signal. As a first step linear regression is applied in which w_s, β_s and w_h, β_h model parameters are learned. For the regression the target variable is the original rating with four ranks, 4 being the best. Then binary thresholds α_s and α_h respectively are learned on the output of the linear regression for the training data. The binary target variable is the thresholded target score: binary target=0 if "score" is [1 or 2] and binary target=1 if "score" is [3 or 4]. The output of the linear regression and its binarized version for each signal is combined using weight vector w_c. The final prediction is obtained by logistic regression with model parameters w and $\bar{\beta}$. All of the model parameters are learned on the training set. The logistic regression input is the output of the ensemble and its target variable is the binarized rating. This final step ensures that we get a calibrated probabilistic output denoted by p.

variable is the rating with four discrete levels [1..4] as described in Section 2.4. As next step the output y_s of the regressor is binarized. This step enables noise supression by learning a threshold α_s. After this stage the outputs $\hat{y_s}$ and $\hat{y_h}$ of the individual signal modelling paths are combined. A weighted average $\bar{y}$ with weight w_c is then calculated. As a final step of the modelling logistic regression is applied to produce calibrated probability p.

Parameters $[w_s, w_h, \beta_s, \beta_h, \alpha_s, \alpha_h, w_c, w, \beta]$ are learned stage-wise in cross validated manner on the training set. The output of the pipeline is a binary prediction and the corresponding probability output from the logistic regression block.

In the static approach the classifier of choice was support vector machine (SVM, [25, 26]) with non-linear radial basis function kernel. After training the decision boundary is represented by "support vectors" which are the most difficult cases from both classes to be distinguished. An advantage of that method is that it can learn complex interactions between features and is not sensitive to class imbalance or skew. A disadvantage is that the required sample size depends on the representation. High ratio of support vectors over sample size indicates that the requirement is not met and the resulting model will have large generalization error on unseen data. In [9] time series were segmented into 10 parts and summary statistics (max, mean, min) were calculated for

each segment. The resulting high dimensional representation was then input to the SVM classifier. In the more recent report of [8] segmentation was dropped and the same summary statistics were calculated over the entire time series of the facial expression estimates (presence of AUs, intensity of given discrete expression, etc.). The resulting representation still had 16 dimensions. We speculate that one of the reasons for the relatively low and variable accuracy was that sample size was too small relative to dimensionality. We opted for a simpler linear ensemble model that has lower dimensionality and captures weak interactions via weighted voting. Once more samples are available our method can be extended to include more features and capture both linear or non-linear interactions between features (generalized stepwise linear regression models can systematically check pairwise or higher order interactions between features).

3. Results and Discussion

We first report test results across all commercials, countries and product categories. We then report results for more fine-grained comparisons. These are models that 1) Include only a single variant for related commercials, which eliminates any bias due to correlation among the sample commercials but may be influenced by the reduced number of commercials; and 2) models that differentiate between product categories and countries. We then compare the current findings that use dynamic features with ones previously reported that use static features. This comparison addresses our hypothesis that dynamic features enable increased accuracy and greater consistency across product categories.

For all comparisons, we report both accuracy and area under the receiver operating characteristics curve (ROC AUC). Accuracy is the sum of true positives and true negatives divided by all cases. It is intuitively appealing but difficult to interpret when distributions are imbalanced. In such cases, accuracy becomes a biased estimator of agreement between a classifier and ground truth [27]. ROC AUC quantifies the continuous relation between true and false positives. If higher rank is assigned to the "positive class" (in our case commercials that scored higher) then the area under the curve gives the probability that a randomly selected positive instance will be ranked higher than a randomly selected negative one. By definition ROC AUC is 0.5 for a random classifier. ROC AUC is unaffected by imbalance between positive and negative cases, although it may mask differences between classifiers in precision and recall [28, 29]). In our data, class imbalance is mild when comparing across product categories and countries (57%), but often is larger when comparing between categories or countries. Thus, accuracy should be interpreted with caution.

To ensure that the trained models do not overfit ([30], in which case models learn to represent noise components in the training data and become unpredictable in new data, we applied different validation schemes to assess generalization capacity of the trained models. Appropriate for the sample size, we used K-fold cross-validation (Kx-CV) in which samples are iteratively split into K disjoint training and test sets and the final performance metrics are averaged over the tests sets. In the tests we used $K = 10$ folds and the procedure was

repeated $n = 10$ times. From the repeated measuresments we calculate confidence intervals at 95% confidence using t-statistics, which is better suited for small sample size. To help interpret the results, we also report a baseline which is a random model with a prior of the class probability of the training data.

As ads can be grouped along model independent factors like regions and product category, particular cross validations can be run where splits are defined by these factors. We will refer to these validation scheme as Leave One Label Out (LOLO) validation. These experiments test robustness of model performance against variations in those factors.

To enable comparison with results in [8] we also conducted Leave One Out (LOO) where test folds contain only one sample. Let us note, however, that for some metrics (ROC AUC in particular) LOO displays strange behaviour when sample sizes become small [31].

We also report results for the case when only one ad variation is selected. While this data filtering may reduce potential ambiguity in the class membership, it reduces sample size, making training more difficult. To avoid any bias induced by arbitrary selections we ran nested cross-validation for ad selection in each group of the ad variations. The reported metrics are then averages over random ad selections.

3.1. Test results on all samples

The dynamic model was trained and cross-validated on all commercials (N=149) without respect to product category or country. ROC AUC was 0.74 with a narrow confidence interval of only ± 0.03 which indicated high reliability. See Table 1.

repeated 10-fold CV	**Accuracy**	**ROC AUC**
Dynamic approach	72.8±2.3%	0.737±0.025
Random baseline	53.7±2.7%	0.50

Table 1: Cross-Validation test on the dynamic approach (dynamic signal + ensemble model) using all sample points. Performance is expressed in Accuracy and ROC AUC. Where appropriate we report confidence interval at 95% confidence as well.

3.2. Robustness against ad variants

When the dynamic model was trained and cross-validated without inclusion of variants (N = 116), ROC AUC decreased by 0.01 and confidence interval decreased from ± 0.03 to ± 0.01 İn this setting we kept only one variation out of several options in each ad group. To counter bias due to random selections we repeat the random ad selection 10 times and run 10-fold CV for each random selection. See table 2.

Results obtained are quite similar to those obtained on all data points. It indicates that in contrast to our original hypothesis about ambiguity in the labels, the ad variations indeed elicit different behavioural responses. In turn, variations can be considered as independent sample.

10-fold CV	Accuracy	ROC AUC
Dynamic approach	72.6±0.8%	0.735±0.009
Random baseline	54.3±1.0%	0.50

Table 2: Cross-Validation test on the dynamic approach (dynamic signal + ensemble model) using random selections of unique variations of the ads.(Sample size N=116). Performance is expressed in Accuracy and ROC AUC. Where appropriate we report confidence interval at 95% confidence as well.

3.3. Robustness against category and country differences

The generalization capacity of the model can be tested by training the model on all but one product category, testing on the one omitted, and then iteratively repeating training and testing for each category. This is referred to as leave-one-label-out cross-validation (LOLO validation). Similarly, the same iterative LOLO can be performed for country.

ROC AUC was nearly identical for all product categories and only slightly lower than what we found when category membership was ignored. The consistency of findings among product categories is remarkable given the variability in the number of samples for each.

ROC AUC for countries was more variable. Highest results were found for France and the US, lowest for Australia and Germany, with those for Russia and UK intermediate.

See table 3 and 4.

Category	Acc.	ROC AUC	#low	#high
Confections	69.6%	0.700	24	22
Food	75.0%	0.700	10	2
Petcare	72.4%	0.694	37	21
Chewing Gum	75.8%	0.702	13	20
average	73.2%	0.699		

Table 3: Generalization performance of the proposed sales prediction model on different product categories. The validation scheme is LOLO so train fold does not contain samples from the category the test ads belong to. #low and #high denote the number of samples in the low and high performing classes, respectively.

3.4. Comparison of Static and Dynamic approaches

The static approach proposed by [8] and our proposed model presented here involved webcam assessments of subjects' responses to the same product categories in four of the same countries. In both cases, sales lift data were provided by MARS, Incorporated. In both cases results were quantified at ROC AUC, but in the static approach only LOO validation was reported, while we reported repeated 10-fold cross-validation. The two major differences between the approaches are the features that represent data and the applied classification model. The two approaches differed in other respects, as well, unrelated to

Region	Acc.	ROC AUC	#low	#high
Australia	59.3%	0.688	18	9
France	86.7%	0.857	8	7
Germany	63.6%	0.677	10	12
Russia	81.8%	0.714	15	7
UK	79.4%	0.759	20	14
USA	79.3%	0.820	13	16
average	75.0%	0.751		

Table 4: Generalization performance of the proposed sales prediction model on ads from different regions. The validation scheme is LOLO so train fold does not contain samples from the region the test ads belong to. #low and #high denote the number of samples in the low and high performing classes, respectively.

types of features, products, or countries. These differences, such as the number of commercials (fewer for the Dynamic model) and the viewing period (more recent and over fewer years for the Dynamic model), and other procedural aspects are unrelated to type of features. In comparing the results for each model, we remain mindful of these other sources of variation.

3.4.1. Statistical analysis

With this caveat in mind, we report the influence of the features on the classification performance. To help the comparison with the past reports on the static approach, the same non linear SVM was trained on the different features. Table 5 reports results for dynamic and static representations as described in Section2. Static features are not exact replicas of the ones used in the [8], but are very similar. Also included are results for head pose and smile dynamics. For the Dynamic model, performance was better when head and face dynamics were combined rather than used exclusively. This suggests that the packaging of nonverbal behavior, head pose and motion, independently contributes to predicting sales lift. For both LOO and 10-fold cross-validation, the dynamic representations produced much higher performance metrics, since using static signal yielded about random chance performance. This finding emphasizes the importance of dynamic features. The magnitude of the difference between Static and Dynamic suggests that procedural differences (such as number of commercials viewed) play at most a minor role. Further research is needed to evaluate this matter. We also report the number of support vectors (#SV) kept after training as an indicator of generalization problems. For 149 samples in 10-fold cross validation scheme, the size of a training fold is about 135. An SVM model cannot generalize well if #SV is as large as the entire training fold. The results confirmed our assumption that low performance of the static features is because classification of high dimensional representations by non-linear SVM requires more data.

While the different classification models performed similarly on the dynamic signals (Dynamic approach (dynamic signal + ensemble model), ROC AUC 0.737 $\pm$0.025, see Table1), the ensemble model of the dynamic approach is

markedly simpler then the obtained SVM models. In turn it is expected to result in smaller generalization error on unseen data. Another advantage is that improvement by adding other behavioural signals increases model complexity in a well controlled way thus preserving generalization of the improved model.

Validation	**Signal**	**ROC AUC**	**#SV**
LOO	Head pose	0.685	127
	Smile dyn.	0.623	107
	Dynamic signals	0.732	122
	Static signals	0.503	130
10-fold CV	Head pose	0.696±0.015	113
	Smile dyn.	66.32±2.15	96
	Dynamic signals	**0.738**±0.018	109
	Static signals	0.580±0.023	118

Table 5: Impact of the static and dynamic signals on the classification performance. The classifier is the same SVM with non-linear radial basis function kernel.

4. Conclusion

One of the biggest challenges in today's market research is the exponential growth of the number of media contents to be analysed since traditional survey based methods do not scale well. In addition, those methods fail to capture the important emotional aspects of the interaction between content and consumers.

We have created a feasible data acquisition system that allows for large scale behavioural data collection and analysis for practical market research. We have also trained a classification model that learned to distinguish ads with high and low sales performance. Although the size and structure of the training data are limited we managed to show that the learned model generalizes well over some factors not used in the modelling. These promising results may pave the way for a new generation of automated, cost-efficient, behavioural cue driven market research tools for analysis.

To further improve methodology, several limitations need to be addressed. Behavioural analysis is based on average responses assuming that individual differences are just random perturbations. However, it is more likely these individual differences carry relevant information about the differences between the ads. Another limitation is that our model does not allow for more complex interactions between observations. Once more samples are available our method can be extended to include more features and it can also capture linear or non-linear interactions between features (generalized stepwise linear regression models can systematically check pairwise or higher order interactions between features). Finally, hybrid models that test conscious recollection and immediate behavioural-emotional responses must be developed to fully understand the impact of ads on consumer behaviour.

5. Acknowledgement

This work was financially supported by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 645094 (SEWA - Automatic Sentiment Analysis in the Wild). The authors would like to thank the Ehrenberg-Bass Institute for Marketing Science for helping in defining the data collection experiment, and MARS, Incorporated for providing the valuable sales lift data that made this research possible.

References

[1] statista.com, U.s. advertising industry - statistics & facts, [Online] (2015).
URL http://www.statista.com/topics/979/advertising-in-the-us/
1

[2] J. Liaukonyte, T. Teixeira, K. C. Wilbur, Television advertising and online shopping, Marketing Science 34 (3) (2015) 311–330. doi:10.1287/mksc.2014.0899.
URL http://dx.doi.org/10.1287/mksc.2014.0899 1

[3] R. B. Zajonc, Feeling and thinking: Preferences need no inferences, American Psychologist 35 (2) (1980) 151–175. doi:10.1037/0003-066x.35.2.151. 1

[4] A. R. Damasio, Descartes' Error: Emotion, Reason, and the Human Brain, 1st Edition, Harper Perennial, 1995. 1

[5] K. R. Scherer, Feelings integrate the central representation of appraisal-driven response organization in emotion, in: A. S. R. Manstead, N. Frijda, A. Fischer (Eds.), Feelings and Emotions, Cambridge University Press, 2004, pp. 136–157, cambridge Books Online. doi:10.1017/CBO9780511806582.009.
URL http://dx.doi.org/10.1017/CBO9780511806582.009 1

[6] D. Hill, Tell me no lies: Using science to connect with consumers, Journal of Interactive Marketing 17 (4) (2003) 61–72. doi:10.1002/dir.10068. 1

[7] D. McDuff, R. el Kaliouby, J. F. Cohn, R. W. Picard, Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads, IEEE Transactions on Affective Computing 6 (3) (2015) 223–235. doi:10.1109/TAFFC.2014.2384198. 1

[8] D. J. McDuff, Crowdsourcing affective responses for predicting media effectiveness, Ph.D. thesis, Massachusetts Institute of Technology Cambridge, MA, USA (2014). 1, 2.3, 2.6, 3, 3.4, 3.4.1

[9] D. McDuff, R. E. Kaliouby, E. Kodra, L. Larguinet, Do emotions in advertising drive sales?, in: Proceedings of ESOMAR Congress, 2013. 1, 2.6

[10] D. Keltner, The signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame, Journal of Personality and Social Psychology (1995) 441–454. 1

[11] Z. Ambadar, J. F. Cohn, L. I. Reed, All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous, Journal of Nonverbal Behavior 33 (1) (2008) 17–34. doi:10.1007/s10919-008-0059-5.
URL http://dx.doi.org/10.1007/s10919-008-0059-5 1

[12] Z. Hammal, J. F. Cohn, C. Heike, M. L. Speltz, What can head and facial movements convey about positive and negative affect?, in: 2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015, Xi'an, China, September 21-24, 2015, 2015, pp. 281–287. doi:10.1109/ACII.2015.7344584.
URL http://dx.doi.org/10.1109/ACII.2015.7344584 1

[13] H. Dibeklioğlu, Z. Hammal, Y. Yang, J. F. Cohn, Multimodal detection of depression in clinical interviews, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ICMI '15, ACM, New York, NY, USA, 2015, pp. 307–310. doi:10.1145/2818346.2820776.
URL http://doi.acm.org/10.1145/2818346.2820776 1

[14] S. Dolnicar, B. Grn, F. Leisch, Quick, simple and reliable : forced binary survey questions, International Journal of Market Research 53 (2) (2011) 231–252. 2.4

[15] M. Vriens, M. Wedel, Z. Sandor, Split-questionnaire designs: a new tool in survey design and panel management, Marketing Research 13 (1) (2001) 14–19. 2.4

[16] J. Orozco, B. Martinez, M. Pantic, Empirical analysis of cascade deformable models for multi-view face detection, Image and Vision Computing 42 (2015) 47–61. doi:10.1016/j.imavis.2015.07.002. 2.5.2

[17] X. Xiong, F. De la Torre, Supervised descent method and its applications to face alignment, in: Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition, 2013. 2.5.2

[18] M. Slaney, A. Stolcke, D. Hakkani-Tür, The relation of eye gaze and face pose: Potential impact on speech recognition, in: Proceedings of the 16th International Conference on Multimodal Interaction, 2014, pp. 144–147. 2.5.3

[19] A. Doshi, M. M. Trivedi, Head and eye gaze dynamics during visual attention shifts in complex environments, Journal of Vision 12 (2) (2012) 9. doi:10.1167/12.2.9. 2.5.3

[20] M. G. Calvo, A. Gutiérrez-García, A. Fernández-Martín, L. Nummenmaa, Recognition of facial expressions of emotion is related to their frequency in everyday life, Journal of Nonverbal Behavior 38 (4) (2014) 549–567. doi:10.1007/s10919-014-0191-3. 2.5.4

[21] M. LaFrance, M. A. Hecht, The social context of nonverbal behavior. Studies in emotion and social interaction, Editions de la Maison des Sciences de l'Homme & Cambridge University Press, 1999, Ch. Option or obligation to smile: The effects of power and gender on facial expression, pp. 45–70. 2.5.4

[22] P. O. Glauner, Deep convolutional neural networks for smile recognition, Master's thesis, Imperial College London, London, UK (2015). 2.5.4

[23] L. Rokach, Ensemble-based classifiers, Artificial Intelligence Review 33 (1) (2009) 1–39. doi:10.1007/s10462-009-9124-7. 2.6

[24] D. Opitz, R. Maclin, Popular ensemble methods: An empirical study, Journal of Artificial Intelligence Research 11 (1999) 169–198. 2.6

[25] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag New York, Inc., New York, NY, USA, 1995. 2.6

[26] C. Cortes, V. N. Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) 273–297. doi:10.1007/BF00994018. 2.6

[27] L. A. Jeni, J. F. Cohn, F. De La Torre, Facing imbalanced data–recommendations for the use of performance metrics, in: Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII '13, IEEE Computer Society, Washington, DC, USA, 2013, pp. 245–251. doi:10.1109/ACII.2013.47.
URL http://dx.doi.org/10.1109/ACII.2013.47 3

[28] T. Fawcett, An introduction to roc analysis, Pattern Recogn. Lett. 27 (8) (2006) 861–874. doi:10.1016/j.patrec.2005.10.010. 3

[29] D. M. Green, J. A. Swets, Signal Detection Theory and Psychophysics, Wiley, New York, 1966. 3

[30] P. I. Good, J. W. Hardin, Common errors in statistics (and how to avoid them), John Wiley & Sons, 2012. 3

[31] A. Airola, T. Pahikkala, W. Waegeman, B. D. Baets, T. Salakoski, A comparison of auc estimators in small-sample studies., in: S. Dzeroski, P. Geurts, J. Rousu (Eds.), MLSB, Vol. 8 of JMLR Proceedings, JMLR.org, 2010, pp. 3–13. 3