Based on Isolated Saliency or Causal Integration? Toward a Better Understanding of Human Annotation Process using Multiple Instance Learning and Sequential Probability Ratio Test

Chi-Chun Lee, Athanasios Katsamanis, Panayiotis G. Georgiou, Shrikanth S. Narayanan

Signal Analysis and Interpretation Laboratory (SAIL), Los Angeles, CA, USA

chiclee@usc.edu, {nkatsam, georgiou, shri}@sipi.usc.edu

Abstract

Human perception is capable of integrating *local* events to generate an overall impression at the *global* level; this is evident in daily life and is utilized repeatedly in behavioral science studies to bring objective measures into studies of human behavior. In this work, we explore two hypotheses considering whether it is the *isolated-saliency* or the *causal-integration* of information that can trigger the global perceptual behavioral ratings as trained annotators engage in tasks of observational coding. We carry out analyses using Multiple Instance Learning and Sequential Probability Ratio Test in a corpus of real and spontaneous distressed couples' interaction with global session-level abstract behavioral coding done by trained human annotators. We present various analyses based on different behavioral detection schemes demonstrating the potential of utilizing these algorithms in bringing insights into the human annotation process. We further show that while annotating behaviors with more positive impression, annotators gather information throughout the session compared to behaviors with more negative impression, where a single salient instance is enough to trigger the final global decision.

Index Terms: multiple instance learning, sequential probability ratio test, behavior annotation, perception, observational coding

1. Introduction

Humans are capable of combining information from multiple perceived *local* events that span over a given time interval to come up with an overall, *global*, description/judgment of often abstract attributes of interest through a complex and integrative internal perception mechanism. This powerful human mechanism has played a major role in aiding research for numerous scientific communities, and is especially relevant in behavioral sciences where human evaluation is repeatedly used as the core methodology for providing grounding evidence in carrying out various analyses. Trained annotators are considered as objective observers providing consistent *global* perceptual ratings on abstract behavioral attributes of interest for the domain experts, after they observe the entire interaction session of the recorded behavioral data [1].

In psychology and psychiatry, studies of comparing behaviors over short, *local*, time scales (a speaking turn or a complete thought unit) versus long, *global*, time scales (an interaction session or a complete clinical trial) have focused largely on the design of an appropriate 'unit' for annotators to carry out behavioral observational coding. The emphasis has been placed mostly on understanding the distinction (pros and cons) between *micro*-analytic and *macro*-analytic behavioral coding standards [2]. In the domain of human perception studies, the Gestalt Principle Theory of Perception [3] - a perception theory for the human visual process - is one such theory linking local structure to global attribute. It states that the human visual perception is holistic (*global*) in nature and is governed by different principles relating to the structures of *local* events. There has not been explicit research work done in analyzing human perception process in the context of high-level behavioral observational coding. In this work, our aim is to bring insights into this *local-global* process using machine learning algorithms within a behavioral detection framework.

We carry out analyses using *Multiple Instance Learning* (MIL) and *Sequential Probability Ratio Test* (SPRT) examining the question of how human annotators give an overall rating of abstract behavioral attributes at an interaction session level. Could it be based on:

1. **isolated saliency** - judging globally based on locally isolated highly informative events ?
2. **causal integration** - judging globally based on integrating information over time in the annotation process ?

MIL presents a probabilistic formulation for extracting highly-salient local events related to the global rating, and SPRT is a statistical formulation that carries the notion of continuously monitoring and aggregating required information for making a decision in a sequential manner.

Most research relies on controlled lab experimentation for studying human perception. In the present paper, we formulate our analysis of human perception as a machine learning and binary detection/classification framework on a large corpus of spontaneous dialogs with multiple human annotations. The analysis is a two-step process that involves studying the extreme human behaviors that are rated consistently by trained annotator (e.g., detection of a high or low degree of *blame*) in the context of distressed couples' interactions. The two-step process is based on the following:

1. identification of *prototypical* local behavioral patterns that are highly-informative about the global human-perception-based ratings
2. detection of extreme global behaviors with derived *prototypical* local behavioral patterns to infer insights about the annotators' perceptual process

We utilize MIL in the first step to discover *prototypical* local behavioral patterns. We assume that these prototypical local behavioral patterns derived from MIL can be perceptually meaningful because of their ability in performing detection of the extreme global behaviors. We then carry out the second step using SPRT-based and saliency-based detection frameworks with these *prototypical* local behaviors. The assumption behind the second step is that the human annotators are trained in a way such that they learn to retain a repertoire of a set of *prototypical* local behaviors, which they utilize internally to decide whether the particular behavior of interest falls into the categories of *high* or *low* degree rating. This second step is formulated to understand the decision mechanism of human annotators as they execute their internal functions of mapping local events to global ratings. We present analysis results with respect to the six different globally-rated behavioral codes (*blame*, *acceptance*, *negative*, *positive*, *humor* and *sadness*) designed for

the purpose of measuring behaviors in conflictual marital interactions. We represent the behaviors at local speaking-turn-level with lexical information computed using *term frequency-inverse document frequency* (tfidf). We present our analyses and discussions using the proposed framework to bring initial insights into the mechanism underlying the annotation process.

The rest of the paper is organized as follows, Section 2 describes research methodology including analysis database, MIL and SPRT framework, and lexical feature representation. Section 3 presents analyses and discussions of our perception analyses. Section 4 describes our conclusion and future work.

2. Research Methodology

2.1. Corpus Description

We carry out our perceptual analysis in the Couple Therapy Corpus [4]. The corpus consists of audio-video recordings and manual word transcripts of severely-distressed couples as they engaged in problem-solving interactions. As a standard practice in behavioral studies, each spouse's behaviors were rated by multiple trained human annotators using expert-designed behavioral coding manuals. Each annotator was instructed to rate each behavioral code (at the *global* session-level) on an integer scale of 1 - 9, where a higher rating indicates the spouse displays more of that behavior, after observing the whole interaction session. We carry out our perceptual analysis on the extreme ratings (25% and 40% of all the available ratings in the corpus: 186 and 280 samples of ratings, respectively) of the six different global codes (*blame*, *acceptance*, *negative*, *positive*, *humor* and *sadness*). Each of the behavioral codes is categorized into *high* and *low* degree of rating.

The rationale behind this selection is two folds. The first is to be consistent with the work done by Katsamanis et al. [5]. Katsamanis et al. utilized the MIL framework, optimized for classification accuracy on the same corpus, to perform binary classification task on the same set of six codes where high accuracies were obtained. The second reason is that the human inter-evaluator agreements are satisfactorily high for these six codes (0.78, 0.75, 0.80, 0.74, 0.76, 0.72) [6], especially for the set of of extreme ratings (very high and very low ratings of these codes). This signifies that not only the annotators' internalization of the code descriptions are consistent but also reduces the confounds of rater variability making our assumption - that the extraction of the *prototypical* local behaviors possessing perceptually-meaningful global behaviors - better justified.

2.2. Multiple Instance Learning

Multiple Instance learning (MIL) is a semi-supervised learning framework when a label, y, is assigned to a bag that consists of multiple unlabeled instances instead of associating a label with every training instance. The original idea of MIL is formulated for a binary classification task ($y \in \{(+1), (-1)\}$). A bag is labeled as $(+1)$-bag if at least one instance in that bag is $(+1)$, and the bag is (-1)-bag if only all instances are (-1). A general way of solving MIL problem is through maximization of *Diverse Density* (DD) function for a feature vector, x, defined as [7]:

$$DD(x) = \prod_{i=1}^{M} \left[\frac{1+y_i}{2} - y_i \prod_{j=1}^{N_i} (1 - e^{-||B_{ij}-x||^2}) \right]$$

where B_{ij} is the j^{th} instance (feature at each speaking turn) of the i^{th} bag (session), N_i corresponds to the number of instances of the i^{th} bag, and M is the total number of bags. Maximizing the $DD(x)$ function can be posed as finding a point (often termed as a concept point, denote as $\mathbf{t}$) in the feature space that is as close to at least one instance from every $(+1)$-bag and as far away from instances in (-1)-bags. The maximization of the DD function with respect to $\mathbf{t}$ can be solved in an expectation-maximization approach (EMDD [8]); we carry out the approach of EMDD in solving the MIL problem in this work.

Using MIL in the DD formulation is intuitively meaningful in the first step of our analysis framework. We use this process to discover two densities (i.e., two concept points) of *prototypical* local behavioral patterns with respect to the global extreme ratings of behaviors (denote as $\mathbf{t}^{(+1)}$ and $\mathbf{t}^{(-1)}$ for the bag of very *high* $(+1)$-bag and very *low* (-1)-bag rating of a behavior, respectively). For each instance, we can compute $P(t^*|B_{ij})$ where $* = \{(+1), (-1)\}$ as measure of whether that instance is close to the $(+1)$ or (-1) concept point using the following:

$$P(t^*|B_{ij}) = \exp(-\sum_k s_k^2 (B_{ijk} - t_k^*)^2) \quad (1)$$

where k indicates the k^{th} feature and s is the scaling vector of the feature. An individual instance's probability is computed for the bag, and the conventional classification decision in EMDD is based on the original formulation of MIL.

2.3. Sequential Probability Ratio Test

Sequential Probability Ratio Test (SPRT) was originally developed in [9] and has since been widely-used for on-line manufacturing quality control and computerized classification test. SPRT, in the context of this paper, can be used to represent the human decision-making process during the annotation, i.e., the human annotator makes a decision over time about whether the behavior each spouse is exhibiting falls in the *high* or *low* side of extreme behaviors as soon as the annotator becomes 'confident' enough after observing a sequence of interaction data. SPRT sequential decision strategy ($S^*_{i,m}$, where m indicate the m^{th} speaking turn in the i^{th} interaction session) given two possible classes/hypotheses $\{(+1), (-1)\}$ can be written as below:

$$S^*_{i,m} = \begin{cases} (+1) & \text{if} \quad LR_{i,m} \geq U^+ \\ (-1) & \text{if} \quad LR_{i,m} \leq L^- \\ continue & \text{if} \quad L^- < LR_{i,m} < U^+ \end{cases}$$

where $\{(+1), (-1)\}$ in this work indicates the extreme *high* and *low* ratings of behavioral codes, and $LR_{i,m}$ is the likelihood ratio defined as below (assuming i.i.d samples):

$$LR_{i,m} = \prod_{j=1}^{m} \frac{P(B_{ij}|t^{(+1)})}{P(B_{ij}|t^{(-1)})} = \prod_{j=1}^{m} \frac{P(t^{(+1)}|B_{ij})P(t^{(+1)})}{P(t^{(-1)}|B_{ij})P(t^{(-1)})}$$

and by assuming uniform prior, $P(t^{(-1)}) = P(t^{(+1)})$,

$$LR_{i,m} = \prod_{j=1}^{m} \frac{P(t^{(+1)}|B_{ij})}{P(t^{(-1)}|B_{ij})} \quad (2)$$

where $P(t^*|B_{ij})$ is given in Equation (1), U^+ and L^- correspond to an upper-bound and a lower-bound confidence threshold computed from user-defined α (Type I error) and β (Type II error). The thresholds are set based on the following guideline according to Wald [9] to simultaneously control for α and β:

$$U^+ = \frac{1-\beta}{\alpha}, \quad L^- = \frac{\beta}{1-\alpha} \quad (3)$$

here, we define $\alpha = \beta = 0.05$. This SPRT formulation can be used to represent a possible 'causal-integration' annotation decision-making process of human annotators.

2.4. Lexical Feature Extraction

We use the same lexical feature extraction as Katsamanis et al. [5] to represent behavioral information of each instance (speaking turn) for training multiple instance learning described in Section 2.2 because lexical information has been shown to use-

Table 1: *Summary of detection results (percentage of accurately detected sessions): numbers in bold indicate the highest performing decision framework for that specific task*

	25% Task (186 samples of ratings)						40% Task (280 samples of ratings)					
	$\text{Salient}_{\text{max}}$	$\text{Salient}_{n\text{Max}}$	SPRT_{st}	SPRT_{con}	SPRT_{res}	SPRT_{hyb}	$\text{Salient}_{\text{max}}$	$\text{Salient}_{n\text{Turns}}$	SPRT_{st}	SPRT_{con}	SPRT_{res}	SPRT_{hyb}
blame	75.3	**79.0**	72.6	72.6	72.0	73.1	**78.2**	77.1	74.6	74.6	73.2	75.7
acceptance	66.7	**74.2**	71.5	68.8	67.7	70.4	68.9	70.7	72.1	72.9	**76.1**	72.1
negative	65.1	73.6	72.6	73.1	**74.2**	72.6	65.7	**69.6**	66.1	66.8	68.2	65.4
positive	69.9	**74.7**	65.1	65.1	62.9	66.1	70.4	**76.1**	67.9	69.3	69.3	69.6
humor	51.6	51.1	50.5	50.5	51.1	50.5	53.8	51.1	50.0	49.6	49.3	49.3
sadness	47.3	47.9	48.4	48.4	47.3	47.8	54.2	50.7	49.6	50.0	50.4	50.4

ful in behavioral prediction task [5, 10]. Lexical information, in this work, is represented by a vector of normalized product of term/word (for the selected set of terms/words) frequencies with inverse document frequencies (tfidf_n) defined as follow:

$$\text{tfidf}_n(t_k|d_j) = \frac{\text{tfidf}(t_k|d_j)}{\sqrt{\sum_{s=1}^{W} \text{tfidf}(t_s|d_j)^2}}$$

where W equals to the number of terms in an instance. $\text{tfidf}(t_k|d_j)$ is computed by counting the number of appearances, n, of every selected term, t_k, in the document, d_j, and appears in D_{t_k} out of the total of D documents using the following:

$$\text{tfidf}(t_k|d_j) = \begin{cases} n \log \frac{D - D_{tk}}{D_{t_k}} & \text{if} \quad D_{t_k} \neq D \\ 0 & \text{if} \quad D_{t_k} = D \end{cases}$$

The selection criterion of terms is based on information gain computed on the training set for each cross validation fold. We choose the terms appearing in the top 0.5% for the task of 25% (186 samples) and the top 1% for the task of 40% (280 samples) as sorted in descending order of information gain.

3. Analyses Results and Discussions

3.1. Analyses Setup

In this work, our goal is to study the possible human annotators' perception process as they engage in tasks of behavioral observational coding. The analyses are based on detection/binary classification framework of extreme behaviors corresponding to the six behavioral codes (*blame*, *acceptance*, *negative*, *positive*, *humor* and *sadness*). The detection evaluation is measured based on 10-fold cross validation (the same couple was restricted to appear in either training or test set only). Lexical feature selection is performed on the training set only, and EMDD is trained using the MIL toolbox [11].

We employ multiple detection schemes emulating different possible annotation decision-making processes. They can be grouped in two major categories: *isolated-saliency* and *causal-integration*. The following is the list of detection schemes and associated descriptions:

Isolated-Saliency:

Saliency likelihood is defined as estimation by computing each instance's, m, likelihood ratio ($LR_{i,m}$) separately without the product using the following form instead of Equation 2:

$$LR^s_{i,m} = \frac{P(t^{(+1)}|B_{im})}{P(t^{(-1)}|B_{im})}$$

- **Salient**$_{\text{max}}$: assign label of $+1$ to the i^{th} session (with a total of l instances) if the max of $LR^s_{i,1\ldots l} > 1$ and vice versa
- **Salient**$_{n\text{Max}}$: same as **Salient**$_{\text{max}}$ except performing majority vote over n largest $|LR^s_{i,j}|$, n is chosen to be 3 in this work empirically

Causal-Integration:

Causal-integration is based on the SPRT framework decision framework described in Section 2.3.

- **SPRT**$_{st}$: standard SPRT as described in Section 2.3, if the algorithm does not terminate before it reaches the end of the session, the label is decided based on the cumulative LR at the last step
- **SPRT**$_{con}$: the same for **SPRT**$_{st}$ except that the detection algorithm does not terminate and continue running through the entire session. A majority vote is performed for the instances that surpass the threshold (U^+, L^- defined in Equation 3) to decide a label for the session
- **SPRT**$_{res}$: the detection algorithm reset cumulative $LR = 0$ whenever it surpasses the pre-defined threshold, and a final majority vote is carried out to decide a final label
- **SPRT**$_{hyb}$: the same for **SPRT**$_{st}$, except that if the algorithm does not terminate before it reaches the end of the session, the label is decided based on the detection scheme, **Salient**$_{n\text{Max}}$

3.2. Detection Results and Discussions

Table 1 summarizes the detection accuracies for the six different detection schemes (Section 3.1) for the six different globally-rated behavioral codes with two different subsets of the data using lexical features (Section 2.4). We present results on both sets of data (25% total, and 40% total), each with equal splits between the two classes. Kastamanis et al. performed the classification with the 25% total. In this work, we focus on interpreting 40% total task as it includes more data, potentially better generalizable interpretations, but present results on both tasks.

The first thing to note is that the overall accuracies reported here are lower compared to [5] despite the fact many of the setups are similar. We think the main differences could be caused by two major reasons: the first is that Katsamanis et al. utilized a variant of MIL framework, which included estimation of multiple modes in DD function (instead of one concept point) along with the second layer of support vector machine. That formulation is beneficial in boosting the overall accuracy although it is rather difficult to apply in sequential decision framework. We decide to use the standard EMDD framework as a starting point in this work because the original formulation provides a more straightforward interpretation and is applicable to be used in SPRT. The second reason is that Katsamanis et al. optimized parameters for the classification accuracies on the test set for the purpose demonstrating an upper-bound of accuracy of the algorithm. Since our assumptions rely on these prototypical local patterns that can carry significant information about the global ratings, we will be focusing our discussion on only the codes for which we obtain a reasonably high accuracy: *blame, acceptance, negative,* and *positive*.

The second observation is that for the *isolated-saliency* detection scheme, taking majority vote on multiple large values of likelihood ratio shows a better and more robust detection scheme compared to taking one single maximum (except for the code, *blame*). Various *SPRT-based* detection schemes that we employ while all demonstrate reasonable accuracies, however, do not show observable trends of differences when comparing among themselves.

The third point to make is that on the overall level, *isolated-saliency* methods seem to obtain a higher accuracy, which may signify, in general, that the salient events can be more informa-

Table 2: *SPRT$_{st}$: median and 75% quantile of decision time, measured as number of turns required divided by the total number of turns of each session for the 40% task*

	accurately-classified		mis-classified	
	median	75% quantile	median	75% quantile
blame	0.20	0.36	0.22	0.43
acceptance	0.15	0.30	0.20	0.40
negative	0.18	0.33	0.21	0.42
positive	0.20	0.43	0.31	0.62
humor	0.25	0.50	0.22	0.47
sadness	0.17	0.29	0.13	0.26

tive in triggering the global ratings from the annotators. However, one of the major strengths of *SPRT-based* methods is that the decision is made in an on-line fashion - the decision can be made quicker while maintaining a fairly high accuracy. Table 2 summarizes the efficiency of the **SPRT**$_{st}$ algorithm. We see that for behavioral codes, *blame, acceptance, negative, and positive*, the SPRT only requires monitoring about the first 15 - 20% and 30 - 43% of speaking turns for 50% (median) and 75% (75% quantile) of the time when the algorithm decides that it is confident enough to make a correct prediction. When the algorithm makes an error, it tends to take longer (20 - 31% and 40 - 62% of speaking turns for 50% and 75% of the time) reflecting the uncertainty/ambiguity in the information for forming the decision. We demonstrated the potential of SPRT$_{st}$ to be used in a real-time monitoring system for the therapists for early signaling on whether the interactions require immediate intervention.

3.3. Isolated-Saliency vs. Causal-Integration

In this section, we present a discussion on compare the accuracies between *isolated-saliency* and *causal-integration* methods with respect to *blame, acceptance, negative,* and *positive* to offer possible insights into question of human annotation process as posed in Section 1.

First, we examine the sessions where both **SPRT**$_{st}$ and **Salient**$_{max}$ make a correct prediction. These sessions constitute 84.5%, 90.1%, 84.2%, and 82.3% (*blame*, *acceptance*, *positive*, and *negative*, respectively) of all the correct decisions made by using **SPRT**$_{st}$. The 75% quantile of the time it takes for the algorithm, **SPRT**$_{st}$, to make a correct prediction is (0.33, 0.37, 0.42, and 0.40) in this overlapping set of sessions. This set of sessions correspond to about 41% - 50% of the total data depending on the behavioral codes, indicating a significant portion of database have the information that can be found in subparts (or as SPRT$_{st}$ intends to do, the beginning) of the session that is relevant for judging global behavioral attributes.

Furthermore, from Table 1, the behavioral codes that have shown major differences between the two categories are *positive*, *acceptance* and *blame* where *isolated-saliency* is better for *positive* and *blame* and *causal-integration* is better for *acceptance* in the 40% task. While there can be confounds because the features (or classifiers) are not optimal, we attempt to interpret based on the trends seen in Table 1 (right). For the code, *blame*, a single salient instance can be highly-indicative (evident in the maximum accuracy obtained in **Salient**$_{max}$, which could mean the use of lexical term at a specific time affects the overall perceptual evaluation of the behavioral code. For behavioral codes that are designed to measure more positive attitude (e.g., *positive* and *acceptance*), the information that affects annotators' decision seems to be more distributed in the session considering **Salient**$_{n\text{Max}}$ and **SPRT**$_{res}$ are the most successful detection mechanisms. This is also in accordance with the established psychological knowledge that negative impression carries more power (saliency) than positive impression [12].

4. Conclusion and Future Works

In this work, through the design of our analysis framework, we show that it is capable of start exploring questions into the human annotation process as to whether it is the *isolated-saliency* that can trigger the final decision or it is based on the *causal-integration* of information. Examining the results, we have demonstrated that not all behaviors can be judged on thin 'slices' (a.k.a., small amounts of data). In cases where these behaviors can be robustly judged with thin slices, these slices need to be contextually appropriate (salient regions). We further reinforce the idea that data dependent modeling of annotator's behavior for automating behavior coding is crucial as in-line with the works [13, 14]. While the results in the work need to be further detailed investigated and verified, it is promising to see that some initial results corroborate the knowledge in psychology.

There are many future directions. One of the limitations in this work is that the use of lexical features computed by tfidf carry only partial information. We should also investigate further the assumption that the notion of perceptually-meaningful local behavioral patterns can be derived from MIL. We plan on incorporating cues from other communicative channels and refining the classifiers within the same conceptual framework to bring further insights into different attentive process on behavioral cues. Lastly, we would like to continue designing the analyses framework for other perceptual experiments and to collaborate with domain experts to further enhance the quantitative aspects toward understanding human judgment of behavior.

5. Acknowledgment

This research was supported in part by the NSF and the Viterbi Research Innovation Fund. Special thanks to the Couple Therapy research staff for collecting, transcribing, and coding the data.

6. References

[1] G. Margolin, P. Oliver, E. Gordis, H. O'Hearn, A. Medina, C. Ghosh, and L. Morland, "The nuts and bolts of behavioral observation of marital and family interaction," *Clinical Child and Family Psychology Review*, vol. 1, no. 4, pp. 195–213, 1998.

[2] K. M. Lindahl, *Methodological issues in family observational research*. Erlbaum, 2001, ch. 2, pp. 23–31.

[3] G. Humhprey, "The psychology of the gestalt," *Journal of Educational Psychology*, vol. 15(7), pp. 401–412, 1924.

[4] A. Christensen, D. Atkins, S. Berns, J. Wheeler, D. H. Baucom, and L. Simpson, "Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples," *J. of Consulting and Clinical Psychology*, vol. 72, pp. 176–191, 2004.

[5] A. Katsamanis, J. Gibson, M. Black, and S. Narayanan, "Multiple instance learning for classification of human behavior observations," in *Affective Computing and Intelligent Interaction*, 2011, pp. 145–154.

[6] M. P. Black, A. Katsamanis, B. R. Baucom, C.-C. Lee, A. C. Lammert, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Toward automating a human behavioral coding system for married couples interactions using speech acoustic features," *Speech Communication*, p. (in press), 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639311001762

[7] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Advances in neural information processing systems*, 1998, pp. 570–576.

[8] Q. Zhang and S. Goldman, "EM-DD: An improved multiple-instance learning technique," *Advances in neural information processing systems*, vol. 2, pp. 1073–1080, 2002.

[9] A. Wald, "Sequential test of statistical hypotheses," *The annals of mathematical statistics*, vol. 16(2), pp. 117–186, 1945.

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

[11] J. Yang, "MILL : A multiple instance learning library," online. [Online]. Available: http://www.cs.cmu.edu/ juny/MILL/index.html

[12] R. F. Baumeister, E. Bratslavsky, C. Finkenauer, and K. D. Vohs, "Bad is stronger than good," *Review of General Psychology*, vol. 5, pp. 323–270, 2001.

[13] A. Kartik and S. S. Narayanan, "Data-dependent evaluator modeling and its application to emotional valence classification from speech," in *Proceedings of Interspeech*, 2010, pp. 2366 – 2369.

[14] ——, "Emotion classification from speech using evaluator reliability-weighted combination of ranked lists," in *ICASSP*, 2011, pp. 4956 – 4959.