A Multimodal Context-based Approach for Distress Assessment

Sayan Ghosh*, Moitreya Chatterjee* and Louis-Philippe Morency
Institute for Creative Technologies, University of Southern California
12015 E. Waterfront Drive, Playa Vista, CA
{sghosh, mchatterjee, morency} @ ict.usc.edu

ABSTRACT
The increasing prevalence of psychological distress disorders, such as depression and post-traumatic stress, necessitates a serious effort to create new tools and technologies to help with their diagnosis and treatment. In recent years, new computational approaches were proposed to objectively analyze patient non-verbal behaviors over the duration of the entire interaction between the patient and the clinician. In this paper, we go beyond non-verbal behaviors and propose a tri-modal approach which integrates verbal behaviors with acoustic and visual behaviors to analyze psychological distress during the course of the dyadic semi-structured interviews. Our approach exploits the advantages of the dyadic nature of these interactions to contextualize the participant responses based on the affective components (intimacy and polarity levels) of the questions. We validate our approach using one of the largest corpus of semi-structured interviews for distress assessment which consists of 154 multimodal dyadic interactions. Our results show significant improvement on distress prediction performance when integrating verbal behaviors with acoustic and visual behaviors. In addition, our analysis shows that contextualizing the responses improves the prediction performance, most significantly with positive and intimate questions.

Categories and Subject Descriptors
G.3 [**Mathematics of Computing**]: Probability and Statistics-dimensionality reduction; I.5.4 [**Computing Methodologies**]: Pattern Recognition-applications; J.3 [**Computer Applications**]: Life and Medical Sciences - health

General Terms
Human Factors, Experimentation, Algorithms

Keywords
Audiovisual Analysis, Distress, Nonverbal Indicators, Psychological Behaviors, Decision Support Technology, Depression, Post-traumatic Stress

*- indicates equal contribution.

ICMI'14, November 12–16, 2014, Istanbul, Turkey

http://dx.doi.org/10.1145/2663204.2663274

1. INTRODUCTION
The prevalence of psychological distress disorders, of the likes of, depression and post-traumatic stress in our society demands a serious effort to create new tools and technologies to help with their diagnosis and cure. This process typically involves a face-to-face dyadic interaction between a clinician and the patient. Recent works in the field have proposed new computational approaches to objectively analyze patient nonverbal behaviors over the duration of the whole session [1, 2, 3]. These techniques have the potential to aid clinicians with their decision for diagnosis or treatment, by giving them a summary of the patient behaviors (i.e., distress indicators) which can be compared with those of the previous sessions of the same person or with a reference population.

Recent approaches in this direction have mostly focused on acoustic, visual and paralinguistic cues for automatically identifying distress indicators [1, 2, 3], ignoring the verbal aspect of the patient responses. Moreover these analyses are performed in a holistic fashion by summarizing the observed cues over the whole interaction. In other words, the responses of the patient are analyzed independent of the context of the questions asked by the interviewers. This discounts the essential information about the affective nature of each stimuli (i.e., questions asked by the clinician/interviewer), which potentially influences the patient's response.

In this paper, we go beyond nonverbal behaviors and propose a tri-modal approach which integrates verbal behaviors with acoustic and visual modalities to analyze psychological distress indicators during dyadic interviews. Our approach takes advantages of the dyadic, semi-structured nature of these interactions to contextualize the participant responses based on the affective components of the questions. In other words, we explore the role of prior knowledge about the affective nature of the stimuli, an individual is subjected to, in predicting psychological distress. Specifically we address this challenge by categorizing the questions asked based on their intimacy and polarity levels. We conduct experiments on a large corpus of 154 semi-structured dyadic interview interactions between a virtual interviewer and a participant.

In the following section, we discuss prior related work in the field of psychology and automatic computational approaches. In Section 3, we present our research hypotheses. Section 4 describes the dataset and the multimodal features, along with our multimodal fusion approach and our experimental methodology. We present the experimental results, along with the feature analysis in Section 5, and conclude the paper with a discussion of future work in Section 6.

2. RELATED WORK

Several researchers in the field of psychology have explored the relationship between both verbal and non-verbal behavior in individuals with general psychological distress and depression. Ellgring has examined the relationship between psychological states and behavior, and its consequences for clinical diagnosis [4]. He investigates the role of non-verbal behavior in depression such as latency in response, motor retardation and lack of emotional variability. Kirsch compared the facial affective behavior of patients suffering from post-traumatic stress disorder with those of healthy subjects, and observed that expressions of joy appear less often in traumatized patients [7]. Fairbanks reported averted eye-gaze, more fidgeting and self-grooming in depressed subjects [8]. Perez and Riggio claim that depressed patients frequently display flattened or negative affect, including less emotional expressivity, heightened anger and gaze aversion [24]. Hall et al. reports shortened speech and lengthened duration of pauses amongst depressed individuals, during verbal interactions [6].

Previous studies have also focused on the automated assessment of psychological disorders. Cohn et al. detected depression by measuring facial actions using AAM (Active Appearance Modeling) and manual FACS (Facial Action Coding System) coding, and prosody using pitch extraction [5]. Stratou et al. explored the role of gender in assessing psychological conditions from recorded video interactions, based on nonverbal behaviors such as affect, expression and motor variability [9]. Valstar et al. suggested looking at both the acoustic and the visual modalities simultaneously [25]. DeVault et al. used paralinguistic cues to investigate the correlation between conversational features and psychological disorders [3]. They conducted their analysis using aggregate dialogue-level features like onset time, filled pauses and speaking rate. Yu et al. proposed a multimodal HCRF (Hidden Conditional Random Field) model to consider commonalities among adjacency pairs of questions to infer psychological states of participants in semi-structured interviews [2].

To our knowledge, this work is the first to propose a context-based computational analysis of psychological distress which integrates verbal behaviors with acoustic and visual. This analysis and integration is performed by taking into account the context of the interviewer questions, obtained by considering the varying degrees of intimacy and polarity of the questions asked.

3. RESEARCH HYPOTHESES

In this section we present the central research hypotheses that we seek to verify through our experiments.

Verbal Behavior: As shown in the previous section, researchers have shown a relationship between para-linguistic cues and psychological distress. For example, research findings report paralinguistic cues such as reduced speech, slow speech, delay in delivery [3, 6]. Inspired by these results and the work of Rude et al who observed, the use of more valenced words amongst people with psychological distress [10], we propose the following hypothesis:

Hypothesis 1a (H1a): Verbal behaviors can be used to predict general distress in individuals.

Integrating verbal and nonverbal information has been shown to improve predictive performance in many social interaction settings. For example, research findings suggested that integrating descriptors of para-verbal behavior with those of non-verbal behavior improve predictability of depressed individuals [2]. This leads us to hypothesize:

Hypothesis 1b (H1b): Integrating markers of verbal behavior with their nonverbal counterparts helps improve the prediction of general distress.

Nature of Stimuli: In the field of psychology, a landmark revelation has been that depressive disorders are manifested by differences in emotional reactivity, such as positive attenuation and negative potentiation. [11, 12]. Thus the affective nature of the stimuli (positive or negative) the participant is subjected to might constitute vital information in terms of predicting psychological distress in humans. We thus propose the following hypothesis:

Hypothesis 2 (H2): Taking into consideration the nature of the affective stimuli favorably influences the task of predicting general distress.

4. EXPERIMENTS

4.1 Dataset

The dataset used in our experiments is an extension of the Virtual Human Distress Assessment Corpus introduced in [26]. It consists of 154 semi-structured interviews between a human participant and a virtual human, an animated human character. Each interaction lasted about 10 minutes on an average. The virtual interviewer, Ellie is controlled in a Wizard of Oz (WoZ) scenario and not only asks questions to the participant, but also provides responses and back-channels, sometimes prompting the participant to expand on a previous answer. This setting requires a human operator sitting behind the wall and deciding on the next spoken utterance of Ellie. The questions asked by Ellie are initially designed to create a rapport with the participant, such as questions about Los Angeles. Subsequent questions are more personal in nature, such as "*Who's someone that has been a positive influence in your life*?" Following this, Ellie switches to questions whose replies may be suggestive of psychological disorders, such as "*How easy is it for you to get a good night's sleep?"*

The participants for the study were recruited via Craigslist and consisted of 183 participants, with 99 males and 84 females. However, due to the errors in logging the data pertaining to certain participants had to be removed. Hence our experimental dataset consists of 154 participants.

Distress Measure Questionnaires are provided to the participants, and the PHQ-9 [16] and PCL-C [15] severity scores are computed by an expert coder based on the questionnaire responses. The severity scores for PHQ-9 gives a measure of depression while that of PCL-C gives a measure of PTSD. In our dataset, about 30% of the participants had a high PCL-C score, while 21% of the participants had high PHQ-9 scores.

It has been observed that the PHQ-9 and PCLC-C scores exhibit significant positive correlation, due to high comorbidity between PTSD and depression [17]. Since we are interested in studying psychological distress in general and developing a decision support tool for healthcare providers, we compute the corresponding z-normalized scores and average them to obtain a measure of general psychological distress, which we refer to as Distress Measure (DM) score. This score is used for computing the ground-truth labels. The ground-truth labels are obtained by

using the median of this DM score as a threshold. Whichever subjects score above the median are considered as positive samples while the rest as negative.

4.2 Question Context

During the course of this semi-structured virtual human interview, the participant is asked a series of questions. They are obtained from a question-bank which was used in [27]. To properly quantify the context around each response of the participant, we propose to analyze two aspects of the interviewer questions: intimacy and polarity. We represent the intimacy level on a scale from 1 (not intimate) to 3 (strongly intimate). The polarity aspect of each question is judged on a Likert Scale defined between -2 (strongly negative) up to +2 (strongly positive), where 0 represents neutral. One big advantage of analyzing interactions with a virtual interviewer is that all questions are asked the same way and the list of questions is well-defined.

The questions were rated by two expert coders for their polarity and intimacy. The ratings were averaged out to determine the intimacy and polarity levels of each question. The inter-coder agreement as measured by Krippendorff's α was 0.86 [14]. This is indicative of a high degree of inter-coder agreement. Each question is either a main question or a follow-up. For the purposes of our experiment all follow-up questions were merged with the main ones. For example, follow-up questions such as "*Can you tell me more*?" are grouped with the previous question, using the same intimacy and polarity label, as noted above.

We propose to categorize the questions in three major groups:

- **Intimate-Positive** This group includes all questions with strong intimacy level (larger than or equal to 1.5) and a positive average polarity score (larger than or equal to 1.5). On an average during the course of each interaction, there were about 3 intimate-positive questions.
- **Intimate-Negative** This group captures intimate questions (larger than or equal to 1.5) with negative polarity scores (less than or equal to -1.5). There were about 7 intimate-negative questions that were asked, during an interaction.
- **Non-intimate** This group represents questions that are not polarized (polarity levels between -0.5 and 0.5) and have low intimacy levels (less than or equal to 0.5). During the course of an interaction, such questions numbered around 11 on an average.

4.3 Multimodal Features

We present in this sub-section the verbal, visual and acoustic features used in our experiments.

Verbal features The textual features have been extracted from transcripts of the participants' conversations with the virtual human. The LIWC (Linguistic Inquiry and Word Count) is a text-analysis program which takes in text as its input, and scans each word in it, finally calculating the normalized term-frequency of the words in each LIWC category [18]. The core of the program is the LIWC dictionary, where each LIWC category (80 in total) is defined based on the social and physiological meaning of words. Words are associated with each category, on the assumption that the categories themselves are linked to social, affective and cognitive processes. They include not only *function* words (such as pronouns, prepositions, articles, auxiliary verbs and conjunctions), but also *emotion* words, which are indicative of positive and negative sentiments. LIWC includes a hierarchical categorization of words such as:

1. Social processes which includes concepts about social partners such as family, friends or, more generally, humans.

2. Affective processes which qualify the emotional state such as anxiety, anger or sadness.

3. Cognitive processes which characterize aspects related to thoughts such as insight, causation and inhibition.

4. Perceptual processes pertaining to the basic senses such as seeing, hearing and feeling.

5. Linguistic processes, which include pronouns like *I, you, we* and assent/negation words such as *yes, OK, no*.

6. Personal concerns relating to issues such as achievements, activities done in leisure, domestic and financial matters.

7. Biological processes which are described by words related to body, health, and sexuality.

These LIWC features have been widely used in the cognitive analysis and study of affect from text and have been applied to different domains, such as prediction of the tie strength in social media [19], and detection of flirtation from speed dates [20].

Visual Features We use visual features obtained from the GAVAM Head Tracker [21], since it has been shown from previous studies [4] that motor variability could be a potential indicator of general distress. GAVAM measures the head rotation in three directions (the pitch, yaw and tilt), along with their means and standard deviations. Five GAVAM features were used in the experiments, corresponding to the mean and standard deviation of pitch and yaw, along with the standard deviation of total rotation in all directions. We also use the CERT (Computer Expression Recognition Toolbox) to measure the Action Units (AUs), which are suggestive of non-verbal expressions [22]. For example, AU 12 corresponds to lip-corner stretching, which is indicative of smiles, and AU 4 corresponds to lowering of eyebrows, suggestive of frowns. CERT also measures the six basic prototypical emotions and expression neutrality such as *Anger*, *Fear*, *Joy*, *Surprise*, *Sadness*, *Contempt*, *Disgust* and expression *neutrality*, which indicates lack of emotions. We use a total of 15 features from CERT, corresponding to six expression based features and nine AU-based features which have been shown to be promising for depression recognition [1].

Acoustic Features For the acoustic modality, we have used 14 acoustic features which have shown promising results in previous studies on psychological disorder analysis. Specifically we use (1) features derived from the glottal source signal obtained by inverse amplitude filtering, such as Normalized Amplitude Quotient (NAQ), Quasi-open Quotient (QOQ) and OQ-NN, a parameter for estimating the open quotient using Mel-frequency cepstral features, and a neural network; (2) H1H2, which is the difference in amplitude (in the spectrum) between harmonics H1 and H2 with low difference for tense voices and high difference for breathy voices; (3) VUV, which is an indicator of whether vocal fold vibration is present and is a measure of the deviation of that vibration (4) Peak slope based features, which identify glottal closure instances from glottal pulses with different closure properties; (5) spectral stationarity for a characterization of prosody range; (6) fundamental frequency for voiced regions of the speech signal; (7) Energy of the speech signal (8) Maxima Dispersion Quotient (MDQ) useful for discriminating breathy and

Table 1. Classification performances of unimodal and multimodal classifiers

Experimental Condition		Precision	Recall	F1-Scores	Accuracy
Unimodal	Text	0.675	0.605	0.6380	63.02%
	Audio	0.622	0.593	0.6071	58.63%
	Video	0.607	0.628	0.6171	58.00%
Audio, Video	Late	0.635	0.628	0.6315	60.50%
	Early	0.642	0.605	0.6227	60.50%
Text, Audio, Video	Late	**0.705**	0.648	**0.6751**	**66.40%**
	Early	0.700	**0.651**	0.6746	66.14%
Majority Baseline		-	-	-	53.86%

tense voices. The interested reader is referred to [17] for a detailed description of the features.

4.4 Prediction Models

In this subsection, we describe the classification models which we use for the automatic assessment of general distress, including an early and a late fusion scheme for combining information from verbal and non-verbal cues. We wish to explore the effect of verbal and non-verbal cues as well as contextual information, thus we have chosen a simple maximum entropy (binary logistic regression) classifier as a basic building block for the models. This ensures that an improvement in the reported classification performances can be attributed to the feature sets, and not to the presence of more sophisticated prediction models.

Majority Baseline As a baseline model, we include a conventional approach where all samples are assigned to the majority label, i.e. with distress. For our experiments, a majority baseline classifier is accurate 53.86% of the time.

Unimodal Classifiers (Verbal, Acoustic or Visual) We used a maximum entropy classifier for each of the individual modalities. The classifier is regularized using an L_2 norm based penalty term, which is validated automatically.

Verbal + Acoustic + Visual (Early Fusion) To fuse information from multiple modalities, such as verbal, acoustic and visual, we use an early fusion scheme where the features from these modalities are stacked together and provided as an input to the classifier (maximum entropy model).

Verbal + Acoustic + Visual (Late Fusion) We employ a probabilistic late-fusion approach to perform a fusion of the features obtained from multiple modalities. For this, we design a two-layered hierarchical model for this task. The first layer consists of three separate maximum entropy classifiers, each trained on the individual modalities. The output probabilities from these classifiers are fused to train a new classifier in the second layer. The classifier in the second layer is trained using a stacked generalization approach [13] and the Expectation-Maximization (EM) algorithm is used to learn the optimal convex weights for combining each of the modalities in the second layer.

Acoustic + Visual (Early and late fusion) As a way to compare with prior work which focused only on nonverbal behaviors, we included two classifiers (early and late fusion) with only the acoustic and visual features.

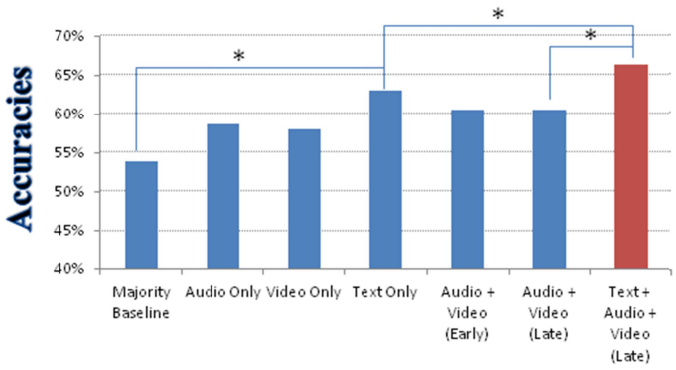

Figure 1: Classification accuracies for implemented models. * indicates statistically significant accuracies with p-value less than or equal to 0.05.

4.5 Methodology

All our experiments follow the Leave-One-Person-Out testing scheme to confirm generalization across participants. Automatic validation of the regularization parameter was performed using a hold-out validation set for each fold. The distress labels were defined by using the median of the distress measure over all participants, as discussed in the Section 4.1. The choice of median as a threshold assured a balanced distribution of the dataset between distressed and non-distressed labels. The experiments have been conducted using scikit-learn [23], a popular open source Python toolbox for machine learning.

We performed automatic feature selection to help with interpretation and performance. Our original set of multimodal features contained 80 verbal features from LIWC, 14 acoustic features from the audio modality, and 20 visual features from the video. We employ the Welch's unpaired t-test to select features with a p-value threshold of 0.10, where the two populations correspond to distressed and non-distressed labels. The t-test takes into account the assumption that the two populations have unequal variances and are normally distributed. A different set of feature was automatically selected for the 4 conditions: No-Context scenario (i.e. computing a single set of features for all the questions asked), Intimate-Positive, Intimate-Negative and Non-Intimate (discussed in Section 4.1), with 23, 13, 17 and 20 selected features respectively.

All predictive models were evaluated with the same standard metrics: precision, recall and F-score, which is the harmonic average of the first two metrics.

5. RESULTS AND DISCUSSION

Our experiments were designed to test our research hypotheses described in Section 3. We first analyzed the effect of verbal behaviors on the task of distress prediction and then studied the role of different context (i.e., affective stimuli).

5.1 Verbal Behaviors and Multimodal Fusion

Our first set of experiments focuses on the efficacy of using verbal behaviors for the prediction of general distress, without any contextual information. Figure 1 and Table 1 show our results comparing unimodal classifiers with a classifier trained by fusing features obtained from the acoustic and visual modalities (acoustic

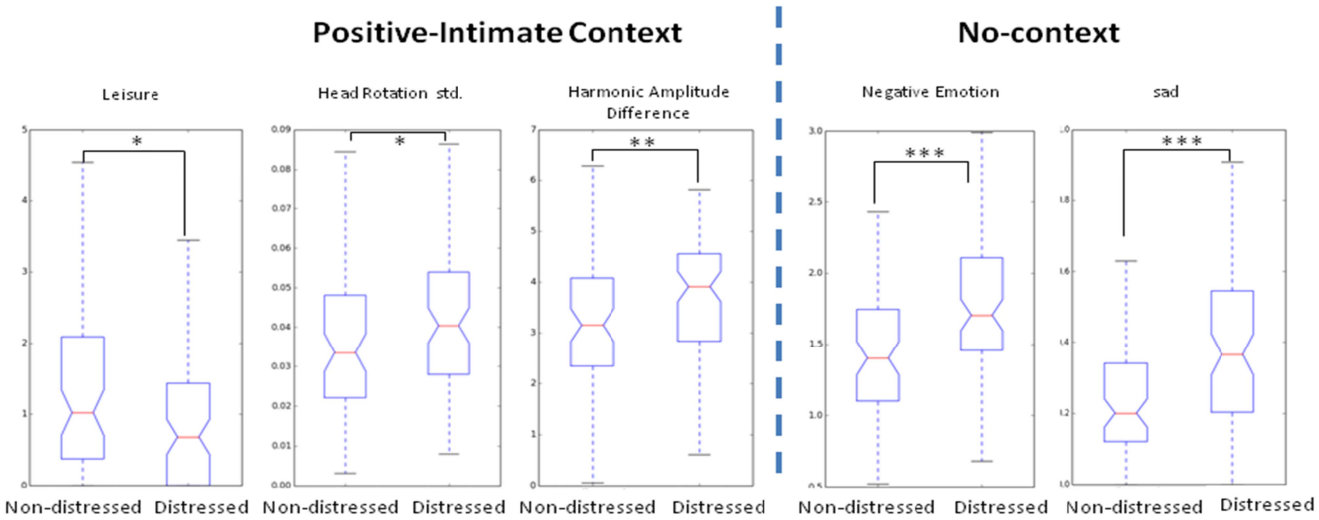

Figure 2: This figure shows the boxplots of some of the statistically significant features for no-context scenario and positive-intimate context. *, ** and ** indicates statistically significant features with p-value <= 0.05, p-value <= 0.01 and p-value <= 0.001 respectively

+ visual). This is followed up by a similar comparison but now with a classifier obtained by fusing the verbal, acoustic and visual modalities (verbal+acoustic+visual), we explore both early and late fusion settings as described in Section 4.3.

Focusing first on Table 1, we observe that the unimodal classifier trained only using verbal features performs marginally significantly better than the majority baseline classifier (see Figure 2), with a t-test statistical result of p=0.05. The direct increase from the majority baseline is 11%.

Here it is worthwhile to point out that other approaches based on multimodal fusion of acoustic and visual features achieved F-scores of 0.664 [2] and 0.88 [1] using more state-of-the-art classifiers such as HCRF or SVM. However the objective of this paper is to explore the discriminative power of the multimodal feature descriptors and hence it is by design that we choose a simpler MaxEnt (Maximum Entropy) model. We also hypothesize that a lower accuracy compared to [1] is also because they used a different dataset (substantially smaller) for a different prediction task, focusing only on predicting depression.

To understand why an improvement in performance is obtained by combining features from the three modalities, we analyze the selected features. Table 2 highlights the most predictable of these selected features for all three modalities, as measured by their p-values, while Figure 2 shows boxplot visualizations for some of them. An analysis of these selected features shows multiple emotionally salient high-level descriptors of verbal behavior to be significant. The significance of descriptors such as *anxiety* and *anger,* which show negative valence, is in concurrence with the findings of [10]. Even verbal markers of positive sentiment such as *assent* or *leisure* are strong predictors. This may be explained by the phenomenon of positive potentiation amongst depressed individuals [11]. Based on these first results, we can confirm hypothesis H1a.

Table 2. A list of the most predictive features under No-Context Condition

Features		p-Values
Text	*Sad*	0.001
	Health	0.001
	negative emotion	0.001
	Anxiety	0.004
	Anger	0.020
	Leisure	0.026
	Negate	0.034
	Hear	0.076
	I	0.085
	assent	0.090
Video	Head Motion std	0.092
	Facial Expression_Neutral	0.085
	Facial Expression_Anger	0.059
Audio	Harmonic Amplitude Diff.	0.012
	Vocal fold vibration deviation	0.022

Furthermore, we also observe from Table 1 and Figure 1 that integrating markers of verbal behaviors with their non-verbal counterparts (visual and acoustic) leads to a statistically significant improvement in the prediction performance, with a p-value=0.04. This result suggests multimodal complementarity when descriptors from the audio (voicing and deviations in amplitude of harmonics) and video (standard deviation of head motion, facial expression of anger) modalities are coupled with their verbal counterparts. This leads to a more accurate prediction over a bimodal late-fusion. Table 2 shows a partial list of selected acoustic and visual features. For example, the visual modality contains negative facial expression which can be complemented with verbal concepts such as anger, negate and sad. This improvement in prediction performance validates our hypothesis H1b.

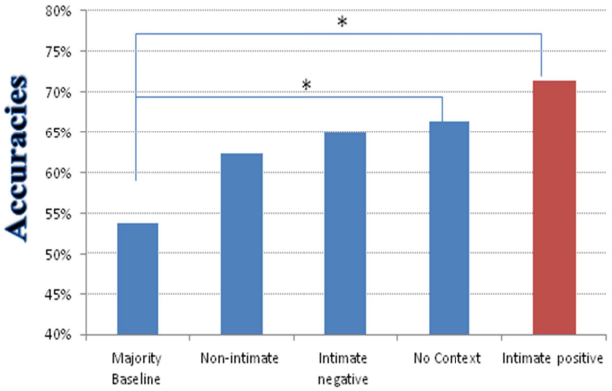

Figure 3: Classification accuracies for various contexts. * indicates statistically significant accuracies with p-value less than or equal to 0.05. All classifiers were trained using tri-modal features with the late-fusion setting.

5.2 Role of Question Context

We designed our second set of experiments to examine the role of the prior knowledge about the affective (positive or negative) nature of stimuli in predicting human psychological distress. As described in Section 4.2, we categorized the interviewer questions in three groups: *Non-Intimate*, *Intimate Positive* or *Intimate Negative*. In the following experiments, we analyze the predictive power of classifiers trained and tested only using the responses to the corresponding specific affective stimuli.

Table 3 and Figure 3 summarize the performances obtained when prior knowledge about the nature of questions asked is known. The results reveal the superiority of the *Intimate Positive* context over the others. This follows from the trend of potentiation of behavioral traits (both verbal and non-verbal) corresponding to positive stimuli amongst distressed individuals as reported in [11]. A relatively lower performance of the other contexts may be explained by a similarity of behavioral traits among individuals of both the distressed and the healthy populace during these scenarios.

Table 4 shows a list of most predictive features for the Intimate-Positive context, as measured by p-value. It is interesting to note that behavioral markers corresponding to both positive and negative affective state of an individual are significant in this context. The positive ones include leisure, achievement in the verbal modality and head nodding (corresponding to the standard deviation of head motion) in the visual modality are particularly interesting since the distressed individuals express a suppressed response on these parameters. The behavioral markers corresponding to the negative affective state of the individual are expressed by the use of words marking negation and the use of sexually abusive words in the verbal modality and facial expression of anger in the visual modality, for instance, hint at a prolonged continuation of a negative affective state of individuals with distress as opposed to ones without it.

Thus our experiments reveal that during the course of the entire interaction between the participant and the interviewer, it is the intimate and positive group of questions that are the most informative. This is in concurrence with our hypothesis H2.

Table 3. Performances of context-based classifiers

Experimental Condition	Precision	Recall	F1-Scores	Accuracy
Intimate Positive	**0.7468**	**0.7108**	**0.7285**	**71.42%**
Non-Intimate	0.6712	0.5903	0.6282	62.33%
Intimate Negative	0.7042	0.6024	0.6451	64.93%

Table 4. A list of most predictive features extracted from the Intimate-positive context

Features		p-Values
Text	*achieve*	0.001
	sexual	0.019
	negate	0.020
	leisure	0.034
	cause	0.069
	you	0.079
Video	Head Motion std	0.035
	Facial Expression_Anger	0.051
	Facial Expression_Joy	0.096
Audio	Harmonic Amplitude Diff.	0.009

It is interesting to note, that the optimal distribution of convex weights (i.e. the weights sum to unity) amongst the three modalities in all scenarios (i.e. intimate-positive, intimate-negative, neutral and non-intimate, no-context) showcases a dominance of the text modality over the others. This is also manifested by the unimodal accuracies reflected in Table 1, which shows the supremacy of the text modality over the others.

6. CONCLUSION AND FUTURE WORK

This paper presented an approach for predicting human psychological distress which integrates verbal, acoustic and visual behaviors. Our results on a large scale dataset of about 160 interactions emphasizes the importance of including verbal behaviors with previously studied acoustic and visual modalities. Our results further highlight the predictive power of using the nature of interview questions, specifically in terms of their intimacy and polarity levels of questions. As future work, we plan to extend this work to include other verbal descriptors such as language model based representations (e.g., unigrams and bigrams) and syntactic information (e.g., part-of-speech tags). Building on top of our significant results with a simple multimodal classifier (maximum entropy model), we are also planning to explore more complex multimodal fusion approaches.

ACKNOWLEDGEMENT

The authors wish to acknowledge Giota Stratou, Prof. Stefan Scherer and Prof. Jonathan Gratch for their insightful suggestions to improve this paper. Sayan Ghosh also acknowledges the USC Viterbi Graduate School Fellowship for funding his graduate studies. The effort here is supported by DARPA under contract W911NF-04-D-0005 and the U.S. Army. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

REFERENCES

[1] Scherer S., Stratou G., and Morency, L.P.. 2013. Audiovisual behavior descriptors for depression assessment. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction* (ICMI '13). ACM, New York, NY, USA, 135-140. DOI=10.1145/2522848.2522886

[2] Yu Z., Scherer S., Devault D., Gratch J., Stratou G., Morency L.P., and Cassell J., 2013. Multimodal Prediction of Psychological Disorder: Learning Verbal and Nonverbal Commonality in Adjacency Pairs, *Workshop on Semantics and Pragmatics of Dialogue (SEMDIAL).*

[3] DeVault, D., Georgila, K., Artstein, R., Morbini, F., Traum, D. 2013. Verbal indicators of psychological distress in interactive dialogue with a virtual human. *Proceedings of SigDial 2013*.

[4] Ellgring, H. 1986.. Nonverbal expression of psychological states in psychiatric patients. *European archives of psychiatry and neurological sciences*, *236*(1), 31-34.

[5] Cohn, J. F., Kruez, T. S., Matthews, I., Yang, Y., Nguyen, M. H., Padilla, M. T, and De la Torre, F. 2009, Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on* (pp. 1-7). IEEE.

[6] Hall, J. A., Harrigan, J. A., and Rosenthal, R. 1996. Nonverbal behavior in clinician—patient interaction. *Applied and Preventive Psychology*, *4*(1), 21-37.

[7] Kirsch, A., and Brunnhuber, S. 2007. Facial expression and experience of emotions in psychodynamic interviews with patients with ptsd in comparison to healthy subjects. *Psychopathology*, *40*(5), 296-302.

[8] Fairbanks, L. A., McGuire, M. T., and Harris, C. J. 1982. Nonverbal interaction of patients and therapists during psychiatric interviews. *Journal of abnormal psychology*, *91*(2), 109.

[9] Stratou, G., Scherer, S., Gratch, J., and Morency, L. P. 2013. Automatic Nonverbal Behavior Indicators of Depression and PTSD: Exploring Gender Differences. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on* (pp. 147-152). IEEE.

[10] Rude, S., Gortner, E. M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, *18*(8), 1121-1133.

[11] Bylsma, L. M., Morris, B. H., and Rottenberg, J. 2008. A meta-analysis of emotional reactivity in major depressive disorder. *Clinical psychology review*,*28*(4), 676-691.

[12] Rottenberg, J., Gross, J. J., and Gotlib, I. H. 2005. Emotion context insensitivity in major depressive disorder. *Journal of abnormal psychology*, *114*(4), 627.

[13] Wolpert, D. H. 1992. Stacked generalization. *Neural networks*, *5*(2), 241-259.

[14] Hayes, A. F., and Krippendorff, K. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*,*1*(1), 77-89.

[15] Blanchard, E. B., Jones-Alexander, J., Buckley, T. C., and Forneris, C. A. 1996. Psychometric properties of the PTSD Checklist (PCL). *Behaviour research and therapy*, *34*(8), 669-673.

[16] Kroenke, K., Spitzer, R. L., and Williams, J. B. 2001. The Phq-9. *Journal of general internal medicine*, *16*(9), 606-613.

[17] Scherer, S., Stratou, G., Mahmoud, M., Boberg, J., Gratch, J., Rizzo, A., and Morency, L. P. 2013. Automatic behavior descriptors for psychological disorder analysis. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on* (pp. 1-8). IEEE.

[18] Tausczik, Y. R., and Pennebaker, J. W. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24-54.

[19] Gilbert, E., and Karahalios, K. 2009. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM.

[20] Ranganath, R., Jurafsky, D., and McFarland, D. 2009. It's not you, it's me: detecting flirting and its misperception in speed-dates. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1* (pp. 334-342). Association for Computational Linguistics.

[21] Morency, L., Whitehill, J., and Movellan, J. 2008. Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on* (pp. 1-8). IEEE.

[22] Bartlett, M., Littlewort, G., Wu, T., and Movellan, J. 2008. Computer expression recognition toolbox. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on* (pp. 1-2). IEEE.

[23] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825-2830.

[24] Perez, J. E., and Riggio, R. E. (2003). Nonverbal social skills and psychopathology. *Nonverbal behavior in clinical settings*, 17-44.

[25] Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., and Pantic, M. 2013. AVEC 2013 – The Continuous Audio/Visual Emotion and Depression Recognition Challenge. In *Proceedings of the 21st ACM International conference onMultimedia 2013* (MM '13).

[26] Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, A., and Morency, L.-P. 2014. The Distress Analysis Interview Corpus of Human and Computer Interviews. In *Proceedings of Language Resources and Evaluation Conference, 2014.*

[27] DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgilla, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., and Morency, L.-P. 2014. SimSensei Kiosk: A Virtiual Human Interviewer for Healthcare Decision Support. In *Proceedings of Autonomous Agents and Multiagent Systems*, 2014.